TECHNICAL REPORT NO. 3

A Survey of the Theory of Selective Information and Some of its Behavioral Applications

by

R. Duncan Luce

BUREAU OF APPLIED SOCIAL RESEARCH
COLUMBIA UNIVERSITY

Behavioral Models Project
(NR 042-115)

TECHNICAL REPORT NO. 8

A Survey of the Theory of Selective Information And Some of its Behavioral Applications

by

R. Duncan Luce

June 1954

CU-10-54-NONR-266(21)-BASR
Bureau of Applied Social Research
New York 27, N.Y.

Table of Contents

A Survey of the Theory of Selective Information and Some of its Behavioral Applications[1]

Part I. The Discrete Theory

1. Introduction

There is a widespread belief - most forcefully articulated by Norbert Wiener [99] - that we are undergoing a new scientific revolution, one comparable in scope and scientific significance to that of the last century; but where the dominant concepts in the previous development were energy, power, and efficiency, the central notions here are information, communication, and feedback. Many current problems stem from attempts to transmit information and to exercise control effectively rather than to achieve an efficient use of energy; little more than chaos would result, for example, were the design of a high-speed computer approached from the energy standpoint. "Information is information, not matter or energy. No materialism which does not admit this can survive at the present day." [p. 155, 99].

What then is information? How is it measured? What scientific statements can be made using the term?

Several schools of thought have developed, each giving its own answers to these questions. In this report we propose to examine the answers

---

1. Most often the title 'information theory' is used without the prefix 'selective'; however, some feel that the simpler title is misleading, especially since there exists a theory of structural information and one of semantic information.

of one of these schools and to indicate some consequences for problems of psychology. But before we turn to this, a certain amount of background material on the history, orientation, and relation of information theory to other theories is appropriate.

It is clear that if Wiener and others are correct in their views, the intuitive concept 'information' must be given at least one precise meaning and maybe more. Considering the variety and vagueness of its meanings in everyday usage, it is an a priori certainty that objections will be raised against any particular formulation, which will surely ignore some of these meanings. This problem - if it be such - has been met many times in science; we need only think of words and concepts like force, energy, work, etc. It is doubtful that a formal definition ever stands or falls because of such debates; it is rather the power and depth of the resulting theory which determines its ultimate fate.

Within the last two decades two distinct attempts have been made to deal with the notion of information, one in Europe, and one in America; these have been complementary rather than competitive. Both theories seem to have arisen from much the same class of applied problems: communication involving electrical signals. The European school, in which the names of Gabor [21, 22, 23, 24, 25, 26] and MacKay [54, 55, 56] are the most important, has been concerned with the problem of the information contained in a representation of a physical situation. As seems intuitively reasonable, the concepts of size and dimensionality are important here. In America, largely as the result of work by Wiener [99, 100] and Shannon [87, 88, 89, 92, 93, 94] a theory of information transmission has developed whose dominant concepts are

those of selection, statistical possibilities, and noise.

In this report we shall not go into a detailed study of the notions of structural and metrical information (the European school), for this theory has had, so far as we have determined, almost no effect on behavioral applications. Of interest to the behaviorist, however, is the apparently overlooked fact that the basic concept of structural information theory is identical to the central assumption of factor analysis. Both theories are concerned with the number of independent dimensions which are required to represent a certain class of data, and the geometrical model of any particular situation is as a point in Euclidean n-space. If we are correct in this observation, it is interesting that basically the same concept has been independently arrived at by both the physicists and the psychologists, and it may be unfortunate that each is unaware of the work of the other.

There are, of course, marked differences of emphasis which reflect the diverse origins and problems. For example, the European information theorists have, in the theory of metrical information, examined in some detail the basic natural units in which the several dimensions can be scaled. Their examples are entirely drawn from physics and so it is not immediately obvious whether any of the scaling work in the behavioral sciences is an independent development of metrical information notions or whether they are totally different. On the other hand, the factor analysts have developed an elaborate matrix machinery suited to the determination of the approximate dimensionality of the Euclidean representation of certain types of data. A comparable machinery

does not appear to exist in structural information theory, though, of course, the close relation of the structural model to matrix theory is apparent.

Our concern, however, is with selective information theory. The central observation of this theory is that for a great many purposes - in particular in the design of communication equipment - one is never concerned with the particular message that is sent but rather with the class of all messages which might be sent and the probability of the occurrence of each. "We are scarcely ever interested in the performance of a communication-engineering machine for a single input. To function adequately it must give a satisfactory performance for a whole class of inputs, and this means a statistically satisfactory performance for the class of inputs which it is statistically expected to receive." [p. 55, 99]. From this point of view, information is transmitted by a selection from among certain alternatives, and the contention is that a selection of an *a priori* rare event conveys more information to the receiver than does one which is expected. This use of 'information' obviously ignores all questions of meaning. "It is important to emphasize, at the start, that we are not concerned with the meaning or the truth of messages; semantics lies outside the scope of mathematical information theory."[1] [p. 383, 7].

---

1. Carnap and Bar-Hillel [6] have presented a theory of semantic information which is based on Carnap's work on inductive logic. Since their approach is different from that of selective information theory, and since, as far as we know, there have been no behavioral applications of it, we have elected not to summarize it here. It may, however, become important, and should therefore not be neglected by the serious student of this area.

It may be useful to introduce at this point three common-sense observations which will be given a precise meaning in the presentation of the theory of selective information - precise to the point where numbers can be attached to them.

1. A person communicating over a noisy telephone line can get less across in a given period of time than he can over a perfectly clear line.

2. Not every letter, nor indeed every word, of a message in any natural language is as important as every other one in getting the sense of the message. For example, the missing letter in 'q_iet' or the missing word in 'many happy _____ of the day' can be filled in, with a high probability of being correct, by anyone knowing English, and therefore in the above context they do not carry much important information.

3. Every person seems to have a limited capacity to assimilate information, and if it is presented to him too rapidly and without adequate repetition, this capacity will be exceeded and communication will break down.

As they stand, it is not immediately obvious that at least some of these statements are not concerned with semantics, or, for that matter, that the whole problem of information transmission is not primarily semantic. One major contribution of information theory is in showing that much of what is implied or suggested in these examples and others like them can be given a precise and useful meaning by a statistical treatment.

We shall delve into this more deeply in the following sections; but first, let us discuss briefly some of the origins of the theory and of the developing interest of behavioral scientists in it.[1] Electrical communica-

---

1. A much more complete history of both the American and European schools has been given by Cherry [7, 8].

tion engineers gradually had been gaining experience in the handling and transmission of information since the early days of the telegraph, telephone, and radio, and during the 1920's this experience began to be formalized as a theory. A most important early paper was that of Hartley [33] in 1928, where the logarithmic measure so characteristic of modern information theory was employed in a simple form. The maturation of the theory, however, resulted from the work of two men, Norbert Wiener of M.I.T. and his former student C.E. Shannon of the Bell Telephone Laboratories. Shannon's papers of 1948 [87, 88] are now the classic formulation of the theory, though the more mathematically inclined reader will find McMillan's recent presentation of the central theorems more satisfactory [63].

Both Wiener and Shannon had much larger interests than improved electrical communication, and they sensed the wider implications of the theory and of several related concepts - feedback being one of the most important. In a series of conferences and seminars dating back to 1941 and continuing to the present, these concepts - sometimes classed under the title of 'Cybernetics,' a word coined by Wiener for this somewhat nebulous discipline - and their applications to the various behavioral sciences have been examined and debated. These meetings[1] have been held largely in the East, many of them in Cambridge, and, as a consequence, the impact of information theory, which has been so strong along the Eastern seaboard, has been less marked in the West.

Many of the empirical sciences dealing with human behavior - psychology, linguistics, physiology, biology, psychophysics, social psychology, neuro-

1. In the introduction to his book Cybernetics [99], Wiener presents a detailed history of the early meetings.

logy, medicine, anthropology - have had representatives at some of these seminars; indeed, these men have organized and dominated many of the meetings. From this there has emerged a small group of analytically inclined behavioral scientists who believe that information theory is, or can be, a useful tool in handling some problems in various of the disciplines. We shall try to indicate some of the uses, and the usefulness, of the theory in the latter half of this report.

Our organization of the material is into two parts. In the first, we shall try to present a motivated synopsis of the discrete theory of selective information. The presentation is most deeply influenced by Shannon's, although there has been some departure from his. In the second part we shall be concerned entirely with applications of the theory to problems in psychology. An attempt has been made to group the papers discussed according to the conventional categories used in psychology. A short summary of Shannon's theory of continuous communication systems appears in an appendix. While this theory is of great importance in electrical application, it has so far been of minor significance in behavioral applications, and so it was felt that it should be separated from the main body of the report.

2. General Concepts

2.1 Communication Systems

Information transmission always occurs within a certain physical framework which in general may be called a communication system. Basically such a system consists of three central parts: a source of messages, a channel

over which the messages flow, and a destination for the messages. The source, which very often is a human being, generates messages (and so information, see section I.2.3) by making a series of decisions among certain alternatives. It is the sequence of such decisions that we call a message in a discrete system. These messages are then sent over the channel, which is nothing more than an appropriate medium which establishes a connection having certain physical characteristics between the source and the destination. Mechanically, this picture is incomplete, since the decisions made by the source must be put into a form which is suitable for transmission over the channel, and the signals coming from the channel must be transformed at the destination into stimuli acceptable to it. Thus, between the source and the channel we introduce a transmitter which serves to "match" the channel to the source, and between the channel and the destination we introduce a receiver which "matches" the channel to the destination. In other words, the transmitter encodes the message for the channel and the receiver decodes it. A schematic diagram of the system is shown in Fig. 1.

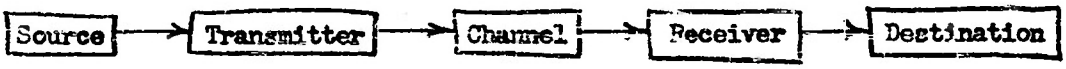

Fig. 1

It is entirely possible to have transmitters which so encode messages that it is not possible to design a receiver which will completely recover the original message. For example, if one has a receiver which encodes all affirmative statements such as "O.K.," "yes," "all right," etc. into

the same signal, then no device can be built which will translate that signal back into the particular word chosen by the source. A transmitter in which this is the case is called singular, otherwise it is called non-singular. (These terms arise if one thinks of the transmitter as a many-many transformation or as a one-to-one transformation.) When the transmitter is non-singular it is possible to design a receiver which is capable of complete recovery of the original message; in other words, there exists a receiver which is the inverse of the transmitter. Throughout our discussion we shall assume that the transmitter is non-singular and that the receiver is its inverse. In effect, this means that we can ignore them in our discussion and suppose that the source and destination are both matched to the channel.

Our abstract communication system seems fairly complete except that it does not allow for the posibility that more than one source may be using the same channel at the same time. Certainly this can happen. It occurs when, by mistake, one telephone line is carrying two conversations at once (crosstalk). It also happens in telephone or radio communication when there is static in addition to the desired message. In all such cases the messages from sources other than the one under consideration - which we will simply call the source - cause interference with messages from the source. Such interference may be minor and may be of no effect on the intelligibility of the message, as for example in the usual low-level telephone static, or it may be most destructive,as when another conversation is cut in. Another example which one might tend to put into the same category of interferences is, say, the 60-cycle hum which is common to so many cheap radios and which is eliminated

in better communication systems only by careful design, a lot of hard work, and a certain amount of luck. If the hum level is high enough it certainly can lower the intelligibility of speech. However, there is an important difference between the problem of interference from hum and that due to static or other conversations. The former is completely predictable, given a short sample to determine the exact frequency, the phase, and the amplitude, and so if it exists in the channel one can either build into the transmitter or into the receiver a network to subtract it from the resulting signal, leaving only the message. Static, hiss, and crosstalk cannot be predicted in any detail from any amount of past evidence about them; therefore, once they enter the channel, they cannot be characterized in full and then subtracted from the signal, but they must rather be accepted and compensated for in other ways.

Thus in our abstraction we must conceive a second source (which may in fact be several lumped together) also feeding signals into the channel, which has the property that (for the problem under consideration) neither the source nor the destination can predict in detail the messages which will emanate from it. The source or the destination may have or may obtain statistical data about the nature of this second source, for example, in an electrical communication system the average power of the second signal may be measured. Such a source is known as a noise source and the signal it generates will be called noise. Clearly, these are often relative terms and what in one context is noise may be the message in another. This, then, completes our model of a

communication system, and it is shown schematically in Fig. 2.

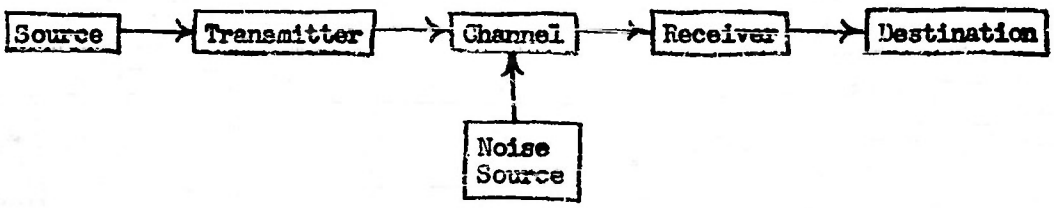

Fig. 2

When there is a noise source in a system it is conventional to speak of the channel as being noisy, but it is well to keep in mind that this is but an abbreviated, and slightly misleading, way of speaking. The noise signal is not an invariant of the channel, as are its physical characteristics. It is clear that one can change the amount of noise in a system while keeping the physical characteristics of the channel, the source, and the destination the same. In any given problem under consideration, the noise level will presumably remain constant and so it can be thought of as a property of the channel, but as we shall see it is a property which must be handled very differently in the theory from the physical characteristics of the channel.

2.2 Noiseless Systems

No communication system is ever noiseless in the sense that there is no noise signal. For example, in any electrical system there must always be

signals present which result from the random agitation of molecules - thermal noise. This can be a serious problem in a high-gain amplifier, but it is not in a telephone. The point, of course, is that noise is not in and of itself bad, but only when it causes a significant interference in the messages sent by the source. The only pertinent feature of noise is whether it causes the destination to 'think' a different message was sent from the one actually sent. Thus if the noise level is low compared with the signal level, so low that it does not significantly alter the message as it passes along the channel, then it may be completely disregarded and the system can be treated as if there were no noise present.

Since we have assumed by definition that the effect of noise is unpredictable in advance except statistically, all we shall be able to state about the effect of noise on messages -- and all we need to state - is the probability that it changes one signal into another. If the signals sent (in a given situation) are always received correctly, then we say the system (or the channel) is noiseless. It must always be kept in mind that if we change the level at which the transmitter operates, or the level of the noise signal, we may change the system from a noiseless one to a noisy one. Being noiseless is a property of the whole system and not of the channel alone!

In principle, it is not necessary to deal separately with the theory of the noiseless and noisy cases, for the former is but a special case of the latter. The presentation, however, is simpler if we bring in the complications one at a time, so we shall examine the noiseless case first (section I.3) and then the noisy one (section I.4).

2.3 The Bit - a Unit of Information

To carry out the program mentioned in the Introduction, namely, to make precise and measurable some of our intuitions concerning the transmission of information, it is necessary to introduce a unit in terms of which amounts of information may be measured. The central observation which is needed before one can arrive at an appropriate unit is that a message conveys information only by its relation to all the other messages which might have been received. Suppose a person is asked whether he smokes. If we have no prior information other than population statistics on smoking, then all we know is the probability that he, as a random selection from the population, will answer 'yes,' and when he selects one of these alternatives and transmits it, some information has been conveyed. But if it is known *a priori*, e.g., from previous conversations or from seeing him smoke that he does smoke, then with probability one the answer will be 'yes' and the receipt of 'yes' from him will not convey any (new) information. In effect, our prior knowledge reduced the set of possible messages to one with but one element, and so far as we are concerned there was no choice to be made, and thus no information could be transmitted.

The minimum condition, therefore, under which information can be transmitted is when there is a choice between two alternatives. The maximum uncertainty in such a choice between two alternatives exists when they are equally probable, and hence the maximum information is conveyed from a choice between two alternatives when they are equally likely. We take such a choice

to be our unit of information. That is, whenever a choice is made between two *a priori* equally likely alternatives (no matter what they are) we shall say that one unit of information has been transmitted by the choice. According to Shannon, Tukey proposed that the unit be called a bit - a shortened form of binary digit - and that term is commonly used. Goldman [29] prefers the term 'bimit' in order to avoid such expressions as 'a bit of information' which, unfortunately, has quite another everyday meaning, but we shall conform to common usage. All of our statements about information transmission, therefore, will be given in this unit; we shall speak of so many 'bits in a message,' or the 'bits transmitted per second,' or the 'bits per English letter,' etc.

A second intuitively desirable feature in measuring information is that it should be additive. We shall formalize exactly what this means later (section I.3.4), but for the present it is enough to say that if two independent choices are made between two *a priori* equally likely alternatives, then a total of two bits is transmitted.

As an example of how the bit may be used, consider a set of elements (think of them as letters of an alphabet) in which each element is equally likely to be selected. Further, suppose that the number n of elements is of the form 2^N where N is an integer. Question: when an element is chosen from this set, how many bits of information are conveyed, i.e., for this set, how many bits per element are there? The answer is N. We can easily show that there are no more than N bits, for suppose we divide the set into half, each

half being composed of 2^{N-1} elements. The element being chosen is in one half or the other , and the decision as to which half it is in is a decision between two equally likely alternatives (since each element has the same probability of being chosen) and so it conveys one bit of information. Take that set and divide it in half, each half now consisting of 2^{N-2} elements. Again, the decision as to which of the two sets contains the desired element is between two equally likely alternatives, and so another bit of information is transmitted in isolating it. Continuing the process until the element is isolated clearly requires N steps, and, assuming additivity, N bits of information are transmitted. The fact that all the elements were assumed to be equally likely should suggest that no scheme can be devised to isolate the element in fewer than N binary decisions; this can be proved to be the case. We shall not prove it, for the conclusion that there are N bits per symbol in this situation will follow from much stronger and deeper results which we shall present later.

The English alphabet consists of 26 letters which with a space, comma, period, semicolon, colon, and question mark totals to $32 = 2^5$ elements. Were we to suppose them to be chosen independently and with equal probabilities (which is patently false) then each letter of a message would yield five bits of information. While this is clearly not a correct estimate of the bits per letter in English prose, it does stand as an upper bound to this number. Later (section II.2) we shall discuss more precise estimates which show that is is actually somewhere between 1 and 2 bits per letter.

Continuing with the example, observe that when $n = 2^H$, then $H = \log_2 n$ by definition of the logarithm, and so we may say that in this situation there are $\log_2 n$ bits of information per element. We will find that our subsequent discussion of information transmission results in logarithmic measures slightly more complicated than this.

3. The Discrete Noiseless System

In this section we shall discuss what is known as the discrete noiseless communication system. The definition of a noiseless system has been given in section I.2.2, and it may be summarized by saying that in such a system there is never any confusion at the destination as to which signal (of a known class of signals) was emitted by the transmitter. This, of course, does not mean that the signal received is necessarily identical to the signal sent, but only that no confusion can arise as to what signal was sent.

The word 'discrete' refers to the nature of the information source, and it describes a source which generates messages by temporally ordered sequences of selections from a finite set of possible choices. Thus, the discrete case includes a vast amount of familiar communication, such as the selections made from an alphabet to generate words and sentences. But the theory of this section does not include sources which can select from all continuous bounded functions on the interval 0 to 1; we have outlined that theory in the appendix.

3.1 Channel Capacity

In any communication system the transmitter is so chosen as to

match the source to the channel. Signals emanating from the transmitter, which are assumed to be in one-to-one correspondence with the selections made by the source, are propagated along the channel. As far as this communication process is concerned, the relevant effect of the physical characteristics of the channel is to determine how many different signals can be transmitted over it in a given space of time. Roughly, this is what we mean by the capacity of the channel. Formally, let $N(T)$ denote the number of different signals which satisfy the following three properties:

i. each signal can be emitted by the transmitter as a result of selections by the source,

ii. each signal is admissible on the channel, i.e., each signal is compatible with the physical characteristics of the channel,

and iii. each signal is of duration T time units.

From the discussion of section I.2.3, it is suggested (though by no means proved) that if each of these $N(T)$ signals were equally likely then there would be $\log_2 N(T)$ bits of information per signal of duration T time units, or

$$C(T) = \frac{\log_2 N(T)}{T}$$

bits per signal per unit time. Now, since it is plausible to suppose that maximum information is transmitted when each signal is equally likely, and since we have taken $N(T)$ to be the largest number of different signals which may be transmitted over the channel in T time units, it is therefore reasonable to suppose that $C(T)$ is approximately the maximum number of bits of information

per signal transmittible over the channel in one time unit. Since there can be only one signal on the channel at a time, C(T) is approximately the maximum number of bits which can be handled by the channel in unit time. The approximation will tend to be better the larger we take T, so we are led to define the capacity C of the channel to be:

$$C = \lim_{T\to\infty} C(T) = \lim_{T\to\infty} \frac{\log_2 N(T)}{T} .$$

For any practical application of this concept the trick is to determine N(T) from the physical characteristics of the channel or from any theorems we may derive which involve C. In the next section we shall discuss any important example of the first procedure, and later, in section I.3.7, we shall present a theorem which has been used to approximate C empirically.

3.2 A Special Case of Channel Capacity

For the moment we shall restrict ourselves to a special class of transmitter-channel combination which, possibly, is best illustrated by the familiar case of the dot-dash telegraphy code. We suppose that at any instant there either is or is not a signal on the wire connecting the transmitter to the receiver. A dot will be represented by one time unit of signal and one time unit of no signal, and a dash by three units of signal followed by one unit of no signal. Between letters we allow three units of no signal and between words six units of no signal. Problem: compute the channel capacity.

Let us define two different states for this system which we shall call a_1 and a_2. The system is in state a_1 following either a letter or a word space, and it is in state a_2 following either a dot or a dash. Since a word or letter space can never follow either a word or letter space, we know that the next signal after the system is in state a_1 must be a dot or a dash, and that the next state must therefore be a_2; however, when the system is in state a_2 it can be followed by any of the four possibilities and so by either state a_1 or a_2. This is illustrated schematically in Fig 3.

We are now in a position to generalize this in a natural manner to a system having m possible states $a_1, a_2, \ldots, a_m$ and n possible signals $S_1, S_2, \ldots, S_n$. When the system is in state a_i then only a certain subset of the signals may arise; let S_s denote a typical one. We suppose that a_i and the admissible S_s together determine what the next state will be. Let us denote it a_j. For all such possible triples (i,s,j) let $b_{ij}^{(s)}$ denote the time duration of the s^{th} symbol. Obviously certain of the combinations cannot arise, e.g., in the telegraphy case the triple (a_1, word space, a_2) is not admissible (see Fig. 3). On the other hand, (a_1, dash, a_2) is admissible and its b value is four time units.

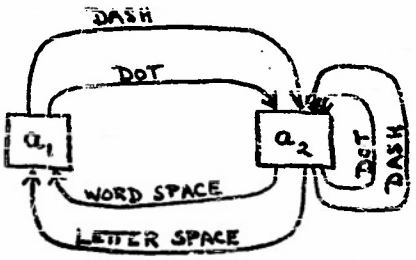

Fig. 3

The channel capacity of this system can be shown [87] to be given

by

$$C = \log_2 W_o,$$

where W_o is the largest real root of the determinantal equation

$$\left| \sum_s W^{-b_{ij}^{(s)}} - \delta_{ij} \right| = 0$$

where

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j . \end{cases}$$

In the telegraphy case, the graph of Fig. 3 can be put in the following matrix form:

		Next State: a_1	Next State: a_2
Present State	a_1	-	dot or dash
	a_2	letter or word space	dot or dash

From this we see that the determinental equation reads

$$\begin{vmatrix} -1 & W^{-2} + W^{-4} \\ W^{-3} + W^{-6} & W^{-2} + W^{-4} - 1 \end{vmatrix} = 0 = \frac{1}{W^{10}} \left[W^{10} - W^8 - W^6 - W^5 - W^3 - W^2 - 1 \right]$$

Solving for W_o and computing $\log_2 W_o$ we find that C = 0.539 bits per unit time.

More will be said about channel capacity before we are done, but first it is necessary to discuss the source and to develop a suitable measure

for the average information generated by any discrete source.

3.3 The Discrete Source

As we have said, we assume that there is a source which makes selections (with replacement) from a finite set of elements and that messages are generated by temporally ordered selections from this set. The situation we have in mind is analogous to the way we form English sentences by ordered selections of letters, blanks, and punctuation marks.

A moment's reflection about English will suggest two important statistical facts about many sources:

i. there is no reason to suppose that the probability that one symbol will be selected is the same as that for another symbol: the letter 's' is much less frequently used in English than is 'e.'

ii. in general, the choice of one symbol in the middle of a message will not be independent of the preceding choices: while 'e' has a high a priori probability of being chosen, the probability is markedly reduced if the letters 'automobi' have already been received and it is markedly increased if the letters 'automobil' have been received.

While most human sources produce an interdependence between symbol selections - often called intersymbol influences - there are some cases of independence, such as the transmission of random numbers or of an unconnected set of telephone numbers. In the next section we shall analyze the case of independent selections and in section I.3.6 the more complicated case where there are dependencies.

To deal with these problems of symbols selected with different

frequencies and of the interdependence of symbol selection, we shall obviously want to introduce probability distributions over the set of symbols. For this to make sense, we shall have to assume that the source is homogeneous in time, so that its statistical character - measured by any statistical parameter we choose - is the same at one time as at any other time. Such a source is said to be stationary and the time series (of symbol selections) is called a stationary time series. This assumption is essential to the theory; it is one which seems plausible for many sources and not for others (we shall return to this in part II on applications). In most cases, however, it is quite difficult to assure oneself that a source is stationary; the problem is very closely related to the difficulty in deciding whether a particular finite set of numbers can be considered a typical sample from a random sequence. The condition serves, however, to prevent us from considering as one source the New York Times from time 0 to time T and Izvestia from time T to time T', for the statistical structure of messages in these two time intervals will certainly be different - indeed, some of the symbols will differ.

Assuming a stationary source S, we may now introduce a little necessary notation. We let $p(i)$ denote the probability that symbol $i \in S$ will be selected and $p(i,j)$ the probability that symbols i and $j \in S$ will be selected in the order i and then j. In general, $p(i,j) \neq p(j,i)$ (consider, for example, q and u in English). In general, if $i_1, i_2, \ldots, i_k$ is an ordered sequence of symbols, $p(i_1, i_2, \ldots, i_k)$ denotes the probability of its occurrence.

The selection of symbols is said to be independent if for every k

and every possible sequence $i_1, i_2, \ldots, i_k$

$$p(i_1, i_2, \ldots, i_k) = p(i_1)p(i_2)\ldots p(i_k).$$

Before turning to the analysis of the case of independent selections, it may be of interest to indicate some of the effects of various assumed statistical dependencies. We shall present the output generated from a source which takes into account some (but not all) of the statistical structure of English. First, suppose that selections are independent but with the simple frequencies of English text. Using these frequencies and a table of random numbers, Shannon [87] generated

OCRO HLI RGWR NMIELWIS EU LL NBNESEBTA TE EKI ALHENRTTPA
OOBTTVA NAR BRL

If, however, one admits intersymbol influence, one may, for example, generate a message in which each selection depends on the two preceding ones. Using such data for English, Shannon generated

IE NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF
DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CHE

Neither message is English, but the second is 'more' English than the first.[1]

3.4 Information Measure for Independent Selections

Let us assume for the present that messages are generated by independent selections from a discrete source. Statistically, then, the source is completely characterised by the probability distribution

$$P = ||p(1), p(2), \ldots, p(n)||$$

---

1. The greater ease the typist found in typing the second passage as against the first is interesting in this connection.

of symbol selection over the n symbols of the source S. The problem is to assign a number to the source, i.e., to the probability distribution P, which we feel is a suitable measure of the average amount of information per symbol in S. There are at least four ways to get to an answer (fortunately the same answer), and since each reveals something of the structure of the problem and since the resulting statistic is of such great importance, we shall present all four.

What we want is a function which assigns a number to each probability distribution; we may denote it by $H = H[p(1), p(2), \ldots, p(n)]$.

The first procedure, which is heuristic and easily remembered, rests on accepting the argument of section I.2.1 that when there are $n = 2^m$ equally likely alternatives, then a suitable measure of the amount of information is $H = \log_2 n$. Let us extend this definition to n equally likely alternatives where n is now any integer, i.e., we shall say there are $\log_2 n$ bits per selection from among n equally likely selections. Now, if we consider any event of probability $p = 1/n$, then we may treat this event as one among n equally likely alternatives and so the information involved in its selection is

$$\log_2 n = \log_2 \frac{1}{p} = -\log_2 p.$$

Finally, consider an event of probability p (not necessarily the reciprocal of an integer): it is plausible to extend the above definitions further and to say that $-\log_2 p$ bits of information are transmitted by the occurrence of this event of probability p. Thus, for the given source S, the selection of symbol i, which occurs with probability p(i), transmits $-\log_2 p(i)$ bits of information.

We see that this has the very reasonable property that an occurrence of a very rare event transmits a great deal of information and an event with probability near 1 transmits almost no information. On the average, however, the amount of information transmitted is the expected value of a single selection from the source, i.e.,

$$H = -\sum_{i=1}^{n} p(i)\log_2 p(i) \quad \text{bits/symbol.}$$

The above expression is without a doubt the best known aspect of information theory, and there are reasons to believe that this expression has blinded some to the content of the theory. It is, of course, nothing more or less than a statistical parameter defined for all distributions and one which is similar to the variance; it obtains meaning and value in only two ways: first, as it is given a meaning in a theory, and second, as it becomes a conventionally accepted way of summarizing certain phenomena. Shannon called H the entropy of the source (or more properly of the distribution characterizing the source) because the same expression arises in statistical mechanics and is called entropy there. There has been considerable controversy as to whether this is only a formal similarity, or whether physical entropy and information are two closely related phenomena. This is a point requiring careful and sophisticated discussion and a rather deeper knowledge of physics than we can assume here. Certain authors have been displeased with the term 'entropy' and they have used terms such as the 'amount of information' or the 'information,' the 'specificity'

and the 'uncertainty of the source.' For the most part, however, 'entropy' is used, and so we shall employ it without committing ourselves to the identity of information and physical entropy.

The second procedure, which many feel to be the simplest and most elegant, amounts to a rigorous formulation of the first one. The technique is to state four intuitively acceptable conditions which must be met by a concept of the information transmitted when a symbol i is selected, given that the a priori probability of its selection was p(i). From these conditions we shall derive the entropy expression; they are:

1. Irrelevancy assumption. The information transmitted by a selection of i shall be a real number which depends only on p(i) and not on the probability distribution over the other symbols. Thus, we may denote the information transmitted by f(p(i)).

2. Continuity assumption. f(p(i)) shall be a continuous function of p(i), for a very small change in p(i) should result in only a small change in the information transmitted.

3. Additivity assumption. If two independent selections i and j with probabilities p(i) and p(j) are effected, then the information transmitted in the joint selection (i,j), which has probability p(i)p(j) of occurring should be the simple sum of the information transmitted by each of the selections, i.e.,

$$f(p(i)p(j)) = f(p(i)) + f(p(j)).$$

4. Scale assumption. In our discussion of the bit, we said that a selection with probability 1/2 shall convey one bit, so we assume

$f(1/2) = 1$.

It follows easily from 3 that $f(p^{\frac{m}{n}}) = \frac{m}{n} f(p)$ for m and n integers, and so by the continuity assumption $f(p^x) = xf(p)$, for any real number x. Any value of p can be written in the form $(1/2)^x$, i.e., as $-\log_2 p = x$. But by 4, $f((1/2)^x) = x$, so $f(p) = f((1/2)^x) = x = -\log_2 p$.

The expected value of the information transmitted by a source with probability distribution p(i) is therefore

$$-\sum_{i=1}^{n} p(i)\log_2 p(i).$$

A third method to obtain the above expression, which is due to Shannon [87], is similar to the last one except that it deals with the whole distribution at once. The procedure is to state a series of four a priori and intuitive conditions which it is felt must be met by any measure of the average amount of information per symbol in the source.

1. The average information transmitted shall be a real-valued function of the n arguments p(1),p(2) ,..., p(n), which we shall denote by H(p(1),p(2) ,..., p(n)).

Next, it seems reasonable, as in the second method, to suppose that if we were to change the distribution very slightly, then H should also change only slightly, so we require that

2. H shall be a continuous function in each of its n arguments.

Further, suppose we consider all sources for which the symbols are equally likely, i.e., $p(i) = 1/n$. As n is increased there is more information transmitted by the selection of one symbol since more messages of a given

length are possible, so we require

3. When $p(i) = 1/n$ for all i, then H is a monotonically increasing function of n.

Finally, we wish to require that if the calculation of the amount of information in a source is divided into a series of subcalculations then the mode of subdivision shall not alter its value. More exactly, suppose S' is a subset of S (which by relabeling we may always take to be the elements $1,2,\ldots,s$). The set S' can, of course, be treated as a single element s' with probability of occurrence

$$p(s') = p(1) + p(2) + \ldots + p(s).$$

If H is known we can compute it for S, for the set with elements s', $s+1$, $\ldots$, n, and for the set S' alone. Our condition asserts that the first number shall be equal to the weighted sum of the last two, i.e.,

4. $$H[p(1), p(2),\ldots,p(n)] = H[p(s'),p(s+1),\ldots,p(n)] + p(s')H\left[\frac{p(1)}{p(s')}, \frac{p(2)}{p(s')},\ldots,\frac{p(s)}{p(s')}\right].$$

From these four conditions, each of which seems to be necessary, Shannon has shown, in a manner not unlike that employed in the second method, that H must be of the form

$$-K \sum_{i=1}^{n} p(i)\log p(i).$$

If we choose the scale constant to be 1 and if we take the logarithm to the base 2, then the binary equally likely case has a value of 1 (as it should to be one bit), and we arrive once again at the entropy expression

$$H = -\Sigma p(i)\log_2 p(i).$$

Before we discuss any of the properties of H and relate it to the other quantity - channel capacity - which we have defined, we shall arrive at the expression for H from the fourth point of view. The following argument is given by Fano [14] and it is similar to one by Shannon [87]. A plausible way to compare two different sources is to define a recoding of any source which takes into account the probability distribution of the source and which results in one of a set of standard normal forms of sources. If these normal forms are such that we can assign a number to each in an intuitively acceptable way, then we have indirectly assigned a number to each source. Of course, the only sources we have associated any numbers to are the binary equally likely ones, so it is more than reasonable that we should attempt a recoding into binary equally likely selections.

This may be done in the following manner. Form all possible messages of length r, i.e., those consisting of r symbols, and call this set R. Since the selections are independent, the probability of each message is simply the product of the probabilities of the individual selections which make it up, hence we know the probability of each message. Thus we have a probability distribution over R. Divide R into a subset R_1 and its complement $\bar{R}_1$ with respect to R in such a manner that the sum of the probabilities of messages in R_1 is as near 1/2 as possible. To each message in R_1 assign the digit 1 and to each in $\bar{R}_1$ the digit 0. Now, divide R_1 into a subset R_2 and its complement $\bar{R}_2$ with respect to R_1 (not R). Again the choice of R_2 is such that the probability of messages in R_2 is as nearly equal as

possible to those in $\bar{R}_2$. To those messages in R_2 assign a second digit 1, so now 11 is assigned to each message in R_2. To those in $\bar{R}_2$ assign as the second digit 0, so 10 is assigned to each message of $\bar{R}_2$. Carry out a similar process in $\bar{R}_1$ leading to the numbers 01 and 00. Continue this 'probability halving' until the classes contain single messages. In this manner each message will have assigned to it a sequence of binary digits, the length of the sequence being in large part determined by the probability of the occurrence of the message - the more probable messages having fewer digits.

An example may make the process clearer:

Message	Probability of occurrence	first digit	second digit	third digit	fourth digit
A	0.50	1	-	-	-
B	0.13	0	1	1	-
C	0.12	0	1	0	-
D	0.12	0	0	1	-
E	0.06	0	0	0	1
F	0.07	0	0	0	0

The first division is between $\{A\}$ and $\{B,C,D,E,F\}$. No further division of $\{A\}$ is possible, and the other set we divided into $\{B,C\}$ and $\{D,E,F\}$. These in turn were divided as $\{B\}$ and $\{C\}$ and as $\{D\}$ and $\{E,F\}$. The final division is of $\{E,F\}$ into $\{F\}$ and $\{F\}$.

Such a coding as this is efficient in the sense that the fewest number of binary digits is assigned to the most probable message and the largest number to the least probable ones. Now, one can ask how many binary digits are required on the average per symbol when messages of length r are

considered. That is, for each message we multiply the number of digits required by the probability that the message occurs, sum these products over all messages, and divide the sum by the total number of symbols r in the message. Call this number H_r. In the above example $H_r = 2.13/r$ bits per symbol. The $\lim_{r\to\infty} H_r$ is a number assigned to each discrete source which both has a plausible meaning and will serve to compare different sources. Fortunately, it can be shown that

$$H = \lim_{r\to\infty} H_r = -\Sigma p(i)\log_2 p(i).$$

Thus by four (really only three) routes we have come to the same statistic as the one which is appropriate to describe the average nature of the source. We can defend it in two further ways; first, by stating some of its properties and by showing that they are reasonable for a measure of information, and second, by using it to make theoretical statements about the transmission of information.

3.5 Properties of H

A number of theorems about H may be proved [07]; as we shall need them later, and as they help to give a feel for H, we shall state them.

i. $H \geq 0$, and $H = 0$ if and only if all p(i) except one equal zero. In other words, the entropy of a distribution is always non-negative, and it is zero if and only if the selection of one symbol is certain. Intuitively, no information is conveyed when the selection is certain, and accordingly $H = 0$.

ii. The maximum value of H is $\log_2 n$ and this maximum is achieved

when and only when each $p(i) = 1/n$.

In words, the maximum average information transmitted per symbol is $\log_2 n$ and that maximum occurs when and only when each of the symbols is equally likely.

It is the above two results which have led many authors to speak of H as the uncertainty of the source, for H has its maximum when what we think of as uncertainty is a maximum and its minimum when absolute certainty obtains.

iii. Let any long message of N symbols be selected and suppose it has probability p of occurring, then $-\log_2 p$ is an estimator of H. This last result is, of course, of considerable importance in estimating H in practical situations, since in general all that can be observed is one message of some long duration. It must be pointed out that when this result is given in precise mathematical language, it asserts that $\frac{\log_2 p}{N}$ almost certainly approaches H as N approaches infinity, i.e., the estimation scheme is consistent.

3.6 Non-independent Selections

So far our discussion of the source has been restricted to the independent case, which, as we pointed out, does not include most sources. But our efforts will not be lost, for fortunately we can readily carry over the results for independent sources to the non-independent case.

We shall consider the selection of one symbol from the set $S = \{1,2,\ldots,n\}$ followed by a second selection (possibly the next one in

forming a message, but we do not need to restrict ourselves to that case). More formally, we let x and y be random variables with range S. The joint distribution of x and y is known and we shall, for convenience, denote the probability that $x = i$ and $y = j$ by $p(i,j)$. In general, of course, $p(i,j) \neq p(i)p(j)$ since the selections need not be independent. The distribution $p(i,j)$ is now defined over the product space SxS, but this differs in notation only from the arbitrary source we have considered earlier, and so our definition of entropy can be applied without alteration to the distribution $p(i,j)$. So we have as the entropy of the joint distribution of x,y,

$$H(x,y) = - \sum_{i,j} p(i,j)\log_2 p(i,j).$$

Similarly, the definition can be applied to the distribution of the random variable x alone and to that of y alone, and so we have

$$H(x) = - \sum_{i,j} p(i,j)\log_2 \sum_{j} p(i,j)$$

$$= - \sum_{i} p(i)\log_2 p(i)$$

and

$$H(y) = - \sum_{i,j} p(i,j)\log_2 \sum_{i} p(i,j)$$

$$= - \sum_{j} p(j)\log_2 p(j)$$

where $p(i) = \sum_{j} p(i,j)$ and $p(j) = \sum_{i} p(i,j)$.

From these definitions Shannon [87] noted the following theorem:

$$H(x,y) \leq H(x) + H(y).$$

This result simply states that the intuitively desirable requirement that the

entropy (or uncertainty or information transmitted) of the joint distribution be no more than the sum of the entropies in the two distributions separately is fulfilled. In addition, Shannon showed that

$H(x,y) = H(x) + H(y)$ if the events x and y are independent. Thus, whenever there is any intersymbol influence in the selections, less information is transmitted per symbol than if they had been independent.

If we introduce the conditional probabilities relating the distribution of y to that of x, further relationships of interest can be established. We let $p(j|i)$ denote the conditional probability that $y = j$ given that $x = i$,

$$p(j|i) = \frac{p(i,j)}{\sum_j p(i,j)} .$$

The conditional entropy of the random variable y given that $x = i$ is defined to be

$$H(y|x = i) = - \sum_j p(j|i)\log_2 p(j|i).$$

Hence the expected conditional entropy of the random variable y given x is

$$H_x(y) = - \sum_i p(i) \sum_j p(j|i)\log_2 p(j|i)$$

$$= - \sum_{i,j} p(i,j)\log_2 p(j|i).$$

$H_x(y)$ measures the average uncertainty in the selection represented by y after the selection denoted by x is known.

Shannon has shown that[1]

---

1. This result is readily proved:

$$H(x) + H_x(y) = - \sum_{i,j} p(i,j)\log_2 \sum_j p(i,j) - \sum_{i,j} p(i,j)\log_2 p(j|i)$$

$$= - \sum_{i,j} p(i,j)\log_2 [\sum_j p(i,j)]\, p(j|i)$$

$$= - \sum_{i,j} p(i,j)\log_2 p(i,j) = H(x,y).$$

$$H(x,y) = H(x) + H_x(y),$$

which, in words, states that the uncertainty of the joint distribution is equal to the uncertainty of the distribution of x added to the uncertainty of that of y when the value of x is known. From this and the preceding result, the following corollary is readily seen to hold:

$$H(y) \geq H_x(y),$$

i.e., the uncertainty of the distribution of y is never increased by a knowledge of x. The two are equal if and only if the two random variables are independent.

One final concept: the ratio of the entropy of a source to the maximum entropy possible with the same set of symbols is a measure of the information transmitting efficiency of the source - Shannon called it the relative entropy. It is generally less than one, either because there is a non-uniform distribution over the symbols or because of the non-independence of symbol selection or, most commonly, because of both. One minus this quantity indicates the percentage of symbols which, though sent, carry no information, i.e., which are redundant. Thus we define the redundancy of a source to be

$$1 - \frac{H}{\max H} = 1 - \frac{H}{\log_2 n} .$$

Several estimation procedures indicate that the redundancy of written English is at least 50 per cent and very likely nearer 75 per cent (see section II.2.1). The reason for such high redundancy will become apparent later.

3.7 The Fundamental Theorem of a Noiseless System

The result we shall state in this section, which is due to Shannon [87], shows in effect that the above definition of channel capacity and of

source entropy or uncertainty are suitable formalizations of our intuitions about the limitations on information transmission.

Theorem: Let the entropy of a source be H bits per symbol and the capacity of a noiseless channel be C bits per second. For any positive number ε, no matter how small, there exists a coding of the output of the source, i.e., there exists a transmitter such that it is possible to transmit at an average rate of $\frac{C}{H} - \varepsilon$ symbols per second. It is not possible to devise a code so as to transmit at an average rate of more than C/H symbols per second.

There are three points which should be made about this theorem. First, it must be kept in mind that the definition of the entropy of a source rests only on the statistical structure of the source, and it does not in any way depend on the properties of the channel. Also, the capacity of the channel depends only on channel properties and not at all on the source. The theorem asserts that these definitions have, however, been so chosen that the ratio C/H is the least upper bound of the transmission rate.

Second, the code which the theorem asserts to exist is, of course, influenced by how small we take ε. If ε is near C/H then nearly any code will do, but as ε approaches 0 fewer and fewer codes will produce a rate of $\frac{C}{H} - \varepsilon$. But the theorem asserts that there will always be at least one. A major unsolved problem of information theory is to devise a theorem which describes such a code in detail for given values of C, H, and ε; the above theorem only asserts that such a code exists.

Third, such optimal use of the channel as described in the theorem

is not effected without paying some price. The price is delay. If one is to code a message optimally when there are intersymbol influences, then it is necessary to wait before transmission to see what that influence is and to make use of it in the coding, thus effecting a delay in the transmission. Similarly, at the receiver, the translation into the language of the destination must be delayed in exactly the same way, for a single received symbol will have meaning only by its relation to a number of others. In practical engineering work a compromise is reached between long delays (and hence expensive storage equipment) and less than optimal use of the channel.

The theorem may be recast in a slightly different form, which may help clarify it and which will be useful when we study the noisy system. Let R denote the average rate at which symbols are transmitted over the channel when a given code is used. The theorem then asserts that $\frac{C}{H} \geq R$ and that there exist codes such that the corresponding R is arbitrarily close to C/H. If we rewrite this as $C \geq HR$ and then maximize both sides with respect to all possible codes we have

$$C = \max_{\text{codes}} C = \max_{\text{codes}} (HR).$$

It is conventional, though misleading, simply to replace HR in the above expression by H. Previously, the entropy of a source was measured in 'bits per symbol,' but for this purpose we measure the entropy of the source (and transmitter combination) in 'bits per symbol' times 'symbols per second,' i.e., in 'bits per second' transmitted. The theorem then asserts that the channel capacity is equal to the maximum number of bits per second which can be transmit-

ted by the source-transmitter combination over the channel. It is this form, and the corresponding form for noisy systems, in which the fundamental theorem has been used in behavioral applications.

4. The Discrete Noisy System

As in the preceding section we shall suppose that the source is discrete, but we shall drop the condition of a noiseless system.

4.1 Equivocation and Channel Capacity

The significant effect of noise in a system, as we pointed out in section I.2.2, is to cause the destination to believe sometimes that a different symbol was transmitted from that which actually was. Any other properties the noise may have are irrelevant to this theory of information transmission. Thus, if we assume that both the signal and the noise time series are stationary, then the noise is completely characterized by the matrix of conditional probabilities $p(j|i)$ which state the probability that symbol j is received when i was sent. Formally, this situation is identical to the case of non-independent selections: in that case we interpreted j as a selection following i, here we shall interpret j as the selection received at the destination when i was actually selected at the source.

The quantities $H(x)$, $H(y)$, $H(x,y)$, and $H_x(y)$ are defined as before. $H(x)$ is the entropy of the source distribution, $H(y)$ the entropy of the destination distribution, $H(x,y)$ the entropy of the joint distribution of x and y, $H_y(x)$ measures the average ambiguity in the signal sent given the received

signal, while $H_x(y)$ measures the average ambiguity of the received signal. When we are considering noise, $H_y(x)$ is called the equivocation.

If a system is noiseless, then $H_x(y) = 0 = H_y(x)$ and so $H(x) = H(y)$.

Let us suppose that all the entropies are calculated in bits/sec, rather than bits/symbol, then the effective average rate of transmission, R, (in bits/sec) is the average rate of information sent, H(x), minus that which was lost as a result of the noise, $H_y(x)$:

$$R = H(x) - H_y(x).$$

This can easily be shown to be equal to two other expressions, the first of which states that the rate of transmission is the difference between what was received and what was received incorrectly. In symbols,

$$R = H(y) - H_x(y)$$
$$= H(x) + H(y) - H(x,y).$$

The notion of rate of transmission for the noisy case is analogous to that introduced for the noiseless case in the last statement of the fundamental theorem of the noiseless case (section I.3.7), and it suggests that one way to define channel capacity in the noisy case is as follows:

$$C = \max_{\text{codes}} [H(x) - H_y(x)].$$

By the theorem of section I.3.7, this definition reduces to that of channel capacity in the noiseless case since $H_y(x) = 0$. However, it does not reduce directly to the definition of channel capacity as given in section I.3.1, but at the end of the next section we shall present a theorem which shows that there is an analogous, though more complicated, definition for the noisy case.

4.2 Theorems

Consider the communication system diagrammed in Fig. 4.

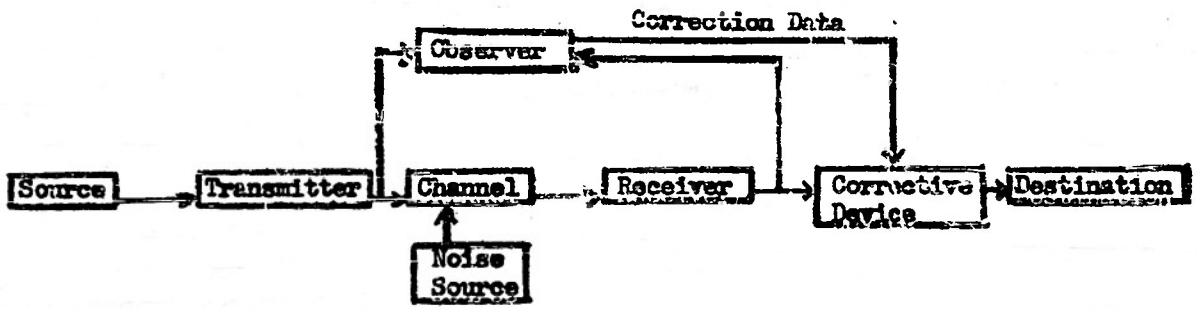

Figure 4

There is assumed to be an observer who is able to perceive both the selections made by the source and the corresponding signals received at the receiver. Let us suppose that the equivocation due to noise is $H_y(x)$, then if there is a noiseless correction channel from the observer to the destination with capacity $H_y(x)$ bits/sec, it can be shown [87] that it is possible to encode correction data in such a manner as to correct all but an arbitrarily small fraction of the errors due to the noise. This is impossible to do if the channel capacity of the correction channel is less than $H_y(x)$. While this theorem is of some theoretical interest, it is certainly not a practical scheme to combat noise. We turn, therefore, to a consideration of communication systems in the sense of section I. 2.1.

The following result, due to Shannon [87], is the fundamental

theorem of the noisy case:

Theorem: Let the entropy of a source be H bits per second and the capacity of the channel C bits per second. If $H \leq C$, then there exists a coding scheme such that the output of the source can be transmitted over the channel with an arbitrarily small frequency of errors. If $H > C$, it is possible to reduce the equivocation to as near $H - C$ as one chooses, but it is not possible to reduce it below $H - C$.

McMillan's comments on this result seem to be worth repetition:

"Engineering experience has been that the presence in the channel of perturbation, noise, in the engineer's language, always degrades the exactitude of transmission. [The theorem] above leads us to expect that this need not always be the case; that perfect transmission can sometimes be achieved in spite of noise. This practical conclusion runs so counter to naive experience that it has been publicly challenged on occasion. What is overlooked by the challengers is, of course, that 'perfect transmission' is here defined quantitatively in terms of the capabilities of the channel or medium, perfection can be possible only when transmission proceeds at a slow enough rate. When it is pointed out that merely by repeating each message sufficiently often one can achieve virtually perfect transmission at a very slow rate, the challenger usually withdraws. In doing so, however, he is again misled, for in most cases the device of repeating messages for accuracy does not by any means exploit the actual capacity of the channel.

"Historically, engineers have always faced the problem of bulk in

their messages, that is, the problem of transmitting rapidly or efficiently in order to make a given facility as useful as possible. The problem of noise has also plagued them, and in many contexts it was realized that some kind of exchange was possible, for example, noise could be eliminated by slower or less 'efficient' transmission. Shannon's theorem has given a general and precise statement of the asymptotic manner in which this exchange takes place." [p. 207, 63]

He goes on to point out the similarity in the exchange between bulk and noise and the rather general exchange between sample size and power in statistical tests.

While the simple repetition of a message is not generally a suitable way to use the channel capacity to eliminate errors, some form of redundant transmission is required. In general it will be far more complicated than repetition, but as with repetition a delay in the reception of a message must result. The essential point of the theorem is that the delay need not be such as to reduce the rate of transmission to zero, as might be thought to be the case. The proof of the theorem is not constructive and so there is no indication of what code to use in order to utilize the channel capacity fully. Shannon writes, "Probably this is no accident but is related to the difficulty of giving an explicit construction for a good approximation to a random sequence." [p. 43, 88] Much recent (engineering) work in information theory has been devoted to finding near optimal codes for certain important special cases.

The fundamental theorem of the noisy case may be recast in a form

which shows the relation of the present definition of capacity to that given for the noiseless case. Let q be a number such that $0 < q < 1$. Consider all possible signals of duration T time units which might be transmitted over the channel and let S denote a typical subset of these signals. Under the assumption that each signal of S is equally probable, let a receiver be designed which is to select from S the most probable element as the cause of the signal it receives. It is clear that in general errors will be made; let p(S) denote the probability that an incorrect interpretation will be made when the subset is S. Consider now all those subsets S such that $p(S) \leq q$. Among these sets there is one which contains the most signals, let that number be denoted by $N(T,q)$. Shannon [87] then showed that

$$C = \lim_{T \to \infty} \frac{\log_2 N(T,q)}{T},$$

which is clearly analogous to the original definition of channel capacity for the noiseless case. It is remarkable that this result is independent of the value of q. Presumably, however, the rate of convergence of the limit is not independent of q, and so in any application of the theorem attempts should be made to exploit the freedom in choosing q.

4.3 Channel Capacity of a Noisy System: Independent Selections

Shannon showed that if one assumes that the selections at the source are independent, then the capacity of the channel is given by the transcendental equation

$$\sum_j 2^{\sum_i h(j|i)\,[C + \sum_j p(j|i)\log_2 p(j|i)]}$$

where $h(j|i)$ is a typical element of the inverse of the noise matrix, i.e.,

$$\sum_j h(j|i)p(j|k) = \delta_{ik}.$$

It is difficult, if not impossible, to see the dependence of channel capacity on the noise matrix from this equation, but, of course, in any given case one can solve numerically for C. However, if we can assume that the noise has the same disturbing effect on each symbol of the source, i.e., if

$$\sum_j p(j|i)\log_2 p(j|i) = \sum_j p(j|k)\log_2 p(j|k)$$

for all i and k, then it can be shown [15] that

$$C = \log_2 n - H(y|x).$$

In the special case of a binary source (two elements) and noise such that the probability of an erroneous transmission is a, then the capacity is given by

$$C = 1 + a \log_2 a + (1-a)\log_2(1-a).$$

It is easy to make interesting calculations using this last expression. For example, if we take the probability of making an error to be 1 per cent, then the channel capacity is reduced to approximately 90 per cent of its value in the absence of noise. This marked non-linearity must be kept in mind whenever thinking about the effects of noise.

5. Some Aspects of Discrete Theory Related to Applications

5.1 Inverse Probabilities, Bayes Theorem, Contingency Tables

As we shall see in some detail in Part II, many of the applications of information theory in psychology are to problems not classically described as communication problems. Indeed, they are communication problems only in the sense that any experiment, or any decision, can be treated as a transmission of information. It is probably more fruitful to remark that in the attempt to analyze communication systems a mathematical formalism has been produced which can be completely divorced from its realization as a communication system. At the same time, there are other realizations of the same mathematical system in psychology. Because of its origins, however, the information terminology is associated with the mathematics and so with any applications which are made. Some of this vocabulary may seem peculiar in the applications, but it is probably not as misleading as it may initially seem. In this section, we propose to discuss, but divorced from the communication model, a part of the formalism which has been particularly important in psychological applications. We shall relate the rate of information transmission to Bayes theorem, we shall generalize the notion of rate of transmission, and we shall discuss the statistical sampling and significance problems.

The structure of very many problems in psychology and in the other behavioral sciences reduces to the existence of two classes of possible occurrences, often called stimuli and responses, such that an occurrence in the response class is in some degree dependent on what stimuli occurred. It is

not easy to characterize in a useful and simple way the relation between these two classes of occurrences. It is, of course, possible to present the whole matrix of joint probabilities p(i,j), i.e., to give the entire contingency table, but this can hardly be called simple. Various measures of contingency have been proposed and used, but objections have been raised to these. Another possibility, and one which certainly has found favor among some psychologists, is the entropy measure. The expression most often used is

$$R = H(x) + H(y) - H(x,y),$$

which, when the entropies are measured in bits/sec, was called the rate of information transmission (section I.4.1). More often than not in the psychological applications time does not enter in a natural manner and it is more appropriate to treat the stimuli and the responses as static and to measure entropies in bits. In that case the following notation is employed:

$$\begin{aligned} T(x;y) &= H(x) + H(y) - H(x,y) \\ &= H(x) - H_y(x) \\ &= H(y) - H_x(y), \end{aligned}$$

and the quantity T(x;y) is simply called the <u>information transmitted</u> from the stimulus to the response. It is a quantity which is 0 when the random variables x and y are statistically independent and it is a maximum when they are in one-to-one correspondence, i.e., when a knowledge of the value of x uniquely determines the value of y. In other words, T is a measure of the contingency between x and y.

Note that in this interpretation of the formalism the role of the human being has changed: Previously, we had thought of the source and the

destination as people and the channel as a physical entity. In most behavioral applications, the stimuli correspond to the source and the responses to the destination; the subject is treated as a noisy channel causing less than perfect correspondence between the stimuli and the responses.

Another way to think about the problem of the relation between the two random variables x and y is in terms of reconstructing the value of x as well as possible from a knowledge of the value of y. This is, of course, the problem of inverse probabilities which has had a long history in statistical theory, and Bayes theorem is one of the most famous results. We may think of it in the following form: There are n possible underlying states of nature, $i = 1,2,\ldots,n$, which are known a priori to have a probability $p(i)$ of occurring. We suppose an experiment is performed with possible outcomes $j = 1,2,\ldots,m$, whose outcome depends somewhat on which state obtains. Let x be a random variable with range the states of nature and distributed according to $p(i)$ and y a random variable with range the experimental outcomes. Further, let us assume as known the conditional probabilities, $p(j|i)$, that $y = j$ when $x = i$. The problem then is to estimate the probability $x = i$ when the outcome of the experiment is known, i.e., when $y = j$ is given.

Cherry describes the analogy to the noisy communication system as "... an observer receives the distorted output signals (the posterior data...) from which he attempts to reconstruct the input signals (the hypotheses), knowing only the language statistics (the prior data)." [p. 39, 8]

It is well known that Bayes theorem reads

$$p(i|j) = \frac{p(j|i)p(i)}{\sum_i p(j|i)p(i)} .$$

If one takes logarithms on both sides of this equation, multiplies the result by $p(i,j)$, and then sums on both i and j, the result is simply

$$H(x) - H_y(x) = H(y) - H_x(y),$$

i.e., the information transmitted from x to y.

5.2 Multivariate Theory

Suppose we are analyzing a stimulus-response situation by information theoretical techniques, then the basic equation we have developed.

$$H(y) = H_x(y) + T(x;y),$$

decomposes an average measure of the response pattern into one part, $T(x;y)$, which is determined by the stimulus plus another, $H_x(y)$, which is unexplained 'random' variation. But it may very well happen that a considerable portion of the residue $H_x(y)$ can be explained in a systematic manner, though not in terms of the experimental stimulus which has so far been considered. For example, consider an experiment in which subjects are required to classify tones which are very near threshold into one of n categories. It may very well happen that the subject's response is only determined in small part by the tone presented, but that in large part it is predictable from a knowledge of his previous response, even if we do not know the stimulus. In such a case, it may be not only appropriate but essential that we consider as the stimulus the pair of random variables (u,v) where u has the possible tones as its range and v the possible previous response of the subject. In other words, in some cases we may be able to understand the phenomena adequately only if we treat as the stimulus a random variable with a range which is the product space of

two, or more, simpler sets. McGill [61, 62] has examined this problem in some detail and he has appropriately generalized the transmission concepts so as to produce a multivariate theory where, of course, Shannon's theory is the bivariate case. We shall recount this development briefly.

First of all, we may replace x by the symbol (u,v), which is equivalent to x when the range of x is the product space of the ranges of the random variables u and v, in the equation for information transmission, and we obtain

$$T(u,v;\, y) = H(u,v) + H(y) - H(u,v;\, y).$$

(We have systematically omitted the extra parentheses about u, v for greater clarity.) It is clear that in our discussion there has been no notion of direction of transmission between source and receiver, and so they may be interchanged, or formally

$$T(u,v;\, y) = T(y;\, u,v).$$

Next, we would like to introduce a measure which gives the separate dependence of y on u and on v. To do this it seems appropriate to define a measure of the conditional information transmitted, which, for example, is the information transmitted from the stimulus u to the response y when the stimulus v is held constant. This, of course, will be an average quantity which in detail is the information transmitted from u to y computed for each possible value of v and then averaged over v. This can be shown to be given by

$$T_v(u;y) = H(v) - H(u,v) - H(v,y) + H(u,v,y).$$

In like manner,

$$T_u(v;y) = H(u) - H(u,v) - H(u,y) + H(u,v,y)$$

$$T_y(u;v) = H(y) - H(u,y) - H(v,y) + H(u,v,y).$$

Clearly, v will have an effect on the transmission from u to y if and only if $T_v(u;v) \neq T(u;y)$, and the magnitude of this effect is measured by

$$A(uvy) = T_v(u;y) - T(u;y).$$

Similar quantities can be defined to measure the effect of u on the transmission from v to y and of y on the transmission from u to v. There is not, however, any need to introduce a new symbol for each of these since they can all be shown to be equal, i.e.,

$$\begin{aligned} A(uvy) &= T_u(v;y) - T(v;y) \\ &= T_y(u;v) - T(u;v). \end{aligned}$$

"In view of this symmetry, we may call A(uvy) the u-v-y interaction information. We see that A(uvy) is the gain (or loss) in sample information transmitted between any two of the variables." [p.5, 62]

With these concepts, it is now possible to express the three-dimensional information transmitted in terms of the two two-dimensional ones and the interaction information:

$$\begin{aligned} T(u,v; y) &= T(u;y) + T(v;y) + A(uvy) \\ &= T_v(u;y) + T_u(v;y) - A(uvy). \end{aligned}$$

We may write this three-dimensional information transmission in another way which parallels the familiar equation $H(y) = H_x(y) + T(x;y)$, namely,

$$\begin{aligned} H(y) &= H_{uv}(y) + T(u,v; y) \\ &= H_{uv}(y) + T(u;y) + T(v;y) + A(uvy). \end{aligned}$$

The term $H_{uv}(y)$ is the residual or unexplained variability in the response y after the information about y given by u and by v and the interaction information of the three variables has been removed.

One unexpected result of McGill's analysis is the possibility that the interaction term may be negative. "In other words, a knowledge of the input [v] may decrease the amount of information that [y] has about [u] - communication from [u] to [y] would actually be better if no data about [v] were collected at all!" [p. 431,72]

One of the most important and desirable properties of the information statistic - entropy - is the additive character, which was apparent in the two-dimensional case and which is even more forcibly illustrated in the three-dimensional theory. Each of the contributions, that from u, from v, from the interaction, and from unexplained variability (while not independent) is simply added to obtain the information in the response pattern. Thus, the information analysis of a stimulus-response situation is somewhat analogous to that of an analysis of variance, and McGill [61] has examined this relation in some detail. But as he pointed out elsewhere, "... information transmission is made to order for contingency tables. Measures of transmitted information are zero when variables are independent in the contingency-sense (as opposed to the restriction to linear independence in analysis of variance). In addition, the analysis is designed for frequency data in discrete categories, while methods based on analysis of variance are not." [pp. 9-10, 62] "It would seem that information theory effectively corresponds to a nonparametric analysis of variance." [p. 431, 72]

There is, naturally, no reason why the above analysis cannot be extended to more dimensions than three, and McGill [62] has carried this out in some detail. There seems little reason to reproduce that here.

In the next section we shall discuss the testing of independence hypotheses in both the multivariate and bivariate cases.

5.3 Statistical Tests and Estimations of Entropy

In addition to the construction of models, behavioral scientists, unlike most physical scientists, must confront the difficult statistical problem of testing and using his model when the only data available are from small samples.[1] His use of information theory is no exception to this rule, so we turn now to that problem.

Let us suppose that a distribution $p(i)$ governs the selections of the k alternatives $1,2,\ldots,k$, and let us suppose that a sample of n independent observations of selections yields $n(i)$ cases of alternative i. The true entropy is, of course,

$$H = -\sum_{i=1}^{k} p(i) \log_2 p(i),$$

while $$H' = -\sum_{i=1}^{k} \frac{n(i)}{n} \log_2 \frac{n(i)}{n}$$ is the estimator of the entropy obtained by replacing each $p(i)$ by its maximum likelihood estimator $\frac{n(i)}{n}$.

Miller and Madow [73] have shown that if the $p(i)$ are not all equal, $\sqrt{n}\,(H-H')$ has a normal limiting distribution with mean 0 and variance

$$\sigma^2 = \sum_{i=1}^{k} p(i) \left[\log_2 p(i) + H\right]^2.$$

If, however, $p(i) = 1/k$ for every i, then $\frac{2n}{\log_2 e}\,(H-H')$ has a chi-square

---

1. In this applied work much calculation is necessary. Newman [74] has described a specialized computer to assist in this. More interesting is the table of $p \log_2 p$ presented by Newman and the more extensive tables of Dolansky and Dolansky [12].

limiting distribution with k-1 degrees of freedom.

They point out that if small samples are used to estimate the entropy there is a bias which can be corrected for by the following theorem:

$$H = EH' + \log_2 e \left[\frac{k-1}{2n} - \frac{1}{12n^2} + \frac{1}{12n^2} \sum_{i=1}^{k} \frac{1}{p(i)}\right] + O\left(\frac{1}{n^3}\right),$$

where EH' is the expected value of H' and $O(\frac{1}{n^3})$ denotes terms of the order of $1/n^3$ or smaller. They also establish a similar expression for the variance of H', but as it is fairly complex we shall not reproduce it here.

For the case of equally likely alternatives, Rogers and Green [85] have developed an exact expression for the expected value of H, namely,

$$EH' = \log_2 n - \sum_{i=2}^{n} \binom{n-1}{i-1} \frac{\sum_{j=0}^{i-2} (-1)^j \binom{i-1}{i-j-1} \log_2 (i-j)}{k^{i-1}}$$

The Miller and Madow approximation in the same case reduces to

$$EH' = \log_2 k - \frac{(\log_2 e)(k^2 + 6n(k-1) - 1)}{12n^2}$$

which, of course, is much simpler. Rogers and Green point out that for $n \geq k$, the two give nearly the same results, but that for $n < k$, "... the Miller-Madow formula ... becomes increasingly less accurate and [their formula] becomes more easily computable." [p.2, 85] They also present a similar expression for the variance which we shall not reproduce here. In another paper [86] they present tables of the mean and variance in the equally likely case for various values of n and k.

Miller [?0] has also treated the problem of contingency tables having r stimulus alternatives and s response alternatives. We let the three probability distributions be $p(i)$, $p(j)$, and $p(i,j)$, and the observed sample frequencies $n(i)$, $n(j)$, and $n(i,j)$ from a sample of size n. The transmitted information, T, is of course given by

$$T = -\sum_{i=1}^{r} p(i) \log_2 p(i) - \sum_{j=1}^{s} p(j) \log_2 p(j) + \sum_{i,j}^{r,s} p(i,j) \log_2 p(i,j)$$

and let T' be the estimator which is obtained by replacing each $p(i)$ by its maximum likelihood estimator $\frac{n(i)}{n}$. If

$$\lambda = \frac{\prod_{i=1}^{r} \left(\frac{n(i)}{n}\right)^{n(i)} \prod_{j=1}^{s} \left(\frac{n(j)}{n}\right)^{n(j)}}{\prod_{i,j}^{r,s} \left(\frac{n(i,j)}{n}\right)^{n(i,j)}}$$

it is known from Wilks' [10] likelihood-ratio test of independence that $-2\log_e \lambda$ has the chi-square distribution with $(r-1)(s-1)$ degrees of freedom. It is not difficult to show

$$n(\log_2 e)T' = -2\log_e \lambda,$$

hence $n(\log_2 e)T'$ has a chi-square distribution with $(r-1)(s-1)$ degrees of freedom when the null hypothesis $T = 0$, i.e., when the stimuli and the responses are independent, is true.

In the same paper, Miller showed that

$$T = ET' - \frac{(r-1)(s-1)}{n\log_2 e},$$

and so it is possible to correct for small sample bias. He suggests that n should be at least 5rs in order to make estimates of the information transmitted. He also showed that

$$\text{var}(T') = \frac{2(r-1)(s-1)}{(n\log_2 e)^2} + \frac{4T}{n\log_2 e}.$$

Fortunately, since we do not know T, we do know that n is generally much larger than T, so the last term can be neglected and the variance is given approximately by the first term.

McGill [62] has extended some of the above results to the multivariate case. First, he observes that:

$$\text{if} \left\{ \begin{array}{l} y \text{ is independent of } (u,v) \\ y \text{ is independent of } v \\ y \text{ is independent of } v \text{ when } u \text{ is held constant} \end{array} \right\} \text{ then } \left\{ \begin{array}{l} T(u,v;y) = 0 \\ T(v;y) = 0 \\ T_u(v;y) = 0 \end{array} \right.$$

The last two conditions each imply

$$T_v(u;y) = T(u;y),$$

or, in words, v is not involved when either of the two conditions holds in the transmission between u and y.

There are, of course, analogous statements for the symbols u, v, and y.

To test the hypothesis that any of the T's are zero, McGill uses Miller's result relating independence with the likelihood-ratio test. One obtains

$$\text{if} \begin{Bmatrix} T(u,v;y) = 0 \\ T(u;y) = 0 \\ T(v;y) = 0 \\ T_y(u;v) = 0 \end{Bmatrix} \text{ then } \begin{Bmatrix} n\log_2 eT'(u,v;y) \\ n\log_2 eT'(u;y) \\ n\log_2 eT'(v;y) \\ n\log_2 eT'(u;v) \end{Bmatrix} \text{ has approximately a chi-square distribution with } \begin{Bmatrix} (UV-1)(Y-1) \\ (U-1)(Y-1) \\ (V-1)(Y-1) \\ Y(U-1)(V-1) \end{Bmatrix} \text{ degrees of freedom}$$

where U, V, and Y are the number of points in the ranges of u, v, and y respectively, and n is the size of the sample.

He shows that if the null hypothesis

$$p(i,j,m) = p(i)p(j)p(m)$$

is true, then $T(u;y)$, $T(v;y)$ and $T_y(u;v)$ are asymptotically independent; thus, as an approximation, the corresponding primed T's can be tested simultaneously for significance under the null hypothesis.

McGill present an interesting example which shows very graphically that "... we cannot decide whether an amount of transmitted information is big or small without knowing its degrees of freedom." [p. 16, 62]

Part II. Applications to Behavioral Problems

1. Introduction

The applications of information theory, including its indirect influences in applied areas, are not easy either to evaluate or to summarize. There can be little doubt that, in addition to the direct applications which we can cite, it has had a very broad impact on the thinking of many behavioral scientists. It has affected both the approach to the analysis of certain types of data and the choice of problems to be considered experimentally. Such influences cannot be succinctly described or tabulated, and we shall not attempt to do so here. A more tangible effect of the theory in the behavioral sciences is the published papers in which it has been explicitly employed. But since these articles have appeared sporadically in most of the behavioral areas, one can hardly hope for a clear pattern of applications.[1] This fact, coupled with the inability of one person to know these various literatures, forces us to consider the two behavioral areas where the publications have been especially numerous and where the pattern is clearer: psychophysics and psychology.[2]

The realization that information theory could play an important role in psychology came in the late forties, only a few years after the

---

1. Biology is to some degree an exception. Much of the application to biological questions has stemmed from the interest of Quastler, who has gathered together much of that work in one volume [8].

2. Much of the material we shall discuss here has been summarized by Miller [72] in somewhat less detail than we shall present here.

publication of Shannon's now classic paper. The realization was symbolized and to a large degree accelerated by a paper which Miller and Frick [65] published in 1949. They observe that "... [a] psychologist's experiments usually generate a sequence of symbols: right and wrong, conditioned and unconditioned, left and right, slow and fast, adient and abient, etc." [p. 314] That is to say, very many experiments are of the stimulus-response type, where the stimuli form one sequence and the responses another. Generally, the procedures to analyze such data have ignored the sequential relations among the responses (usually, though not always, sequential effects in the stimuli have been experimentally eliminated by randomizing procedures), but ignoring the sequential information, they pointed out, is equivalent to assuming the independance of successive responses. It was not implied that psychologists felt that this was a reasonable assumption, but only that the standard statistical techniques were not suited to such an analysis. An exception to this, of course, has been the use of contingency tables to study temporally ordered pairs of responses (digrams) and the use of contingency measures to characterize the degree of association between the arguments of the table. Miller and Frick then outlined certain aspects of information theory and proposed that the information measure be employed in such situations. As Frick and Klaesmar point out in a later paper, "The [information] measure may be applied without logical difficulty to any situation in which one is willing to identify the members of the stimulus and response classes and make some statements about their probability distributions. Whether or not the

measure is useful in the analysis of human behavior remains to be proven. Early results from its application are, however, encouraging..." [p. 15, 19].

There are difficulties, however, for as Miller and Frick pointed out, there are two serious limitations on the applicability of information theory:

1. Sequential responses which are generated while learning is occurring do not form a suitable sample from which to estimate the probabilities which are needed, for the assumptions of learning and of a stationary response time series are incompatible.

2. The difficulty of obtaining adequate samples to estimate probabilities increases sharply with an increase in the length of dependencies in the response sequence; in fact, beyond three step dependencies it is completely out of hand.

Related to the last point are the computational difficulties which arise with large amounts of sequential data. Basically, however, this problem is less serious than the sampling one, since computation machines ideally suited to repetitious calculations are available. In addition, special equipment, such as that described by Newman [74], can be constructed to carry out information-type analysis.

Miller and Frick proposed that the quantity which is called redundancy is communication problems (section I.3.6) be called the index of behavioral stereotypy in behavioral applications. It will be recalled that this is defined as

$$1 - \frac{H}{\max H} .$$

It is a quantity which is 1 when the behavior is completely stereotypic and 0 when each of the several alternatives arises with equal probability. A value of k for the index means that, on the average, k per cent of the responses are completely determined and the remainder, 1-k, are maximally uncertain.

Year by year, following the publication of this paper, there has been an increase in the number of papers in psychological journals employing information theory, with almost a flood in 1953. It is not plausible to suppose that this trend will decrease rapidly, if at all, in the next year or two, and so we can be sure that any summary we attempt here will be out of date before it can be very widely read. Yet already there seems to be some pattern to the publications, and so a summary may serve some function, as long as it is kept in mind that it is a cross section of an incompleted trend.

From our knowledge of the theory, it seems reasonable to class the applications in three categories: 1) Those which employ information theory to deal with sequential data, as proposed by Miller and Frick. Sections II.2 and II.8 are illustrative of this approach. 2) Those which employ the formalism of noisy communication (discussed in section I.5) to cope with problems where stimulus and response are not perfectly correlated, e.g., where there are errors of some type. Sections II.5 and II.7 are typical. 3) Those which employ the central theorems of information theory concerning rate of transmission and capacity. Section II.4, and to some extent section II.5, exemplifies this approach.

In anticipation of our survey, three features of the trend of application seem worthy of note. 1) As Miller and Frick suggested would be the case, and as we mentioned in section I.5.1, few of the applications are to problems usually classified as communication problems. 2) The applications do not generally employ the fundamental theorem relating channel capacity and the statistical structure of the source. For example, we know of only two limited attempts to characterize the capacity of a behavioral system other than by observations of the actual rate of transmission. 3) The theory has not generated new problems to be studied in psychology, but rather it has caused researchers to re-examine old problems from a new point of view. In some cases (see section II.5) it has permitted several apparently disparate effects to be included in a single theoretical framework.

The fact that old problems are being considered again does not, unfortunately, mean that new data are not needed. A published experiment rarely fulfils exactly the conditions another worker would like, and, more important, the isolation of sequential dependencies requires a new analysis of the raw data, and it is very rare indeed to find extensive publications of raw data.

2. The Entropy of Printed English

A problem which has intrigued a number of authors, including Shannon, is the estimation of the entropy of printed English (or any other language, for that matter), i.e., the estimation of the average number of bits per letter in a written passage. Put another way, the problem is to characterize the

average sequential dependencies in the written language. If we assume - as may be approximately true - that the English in one book or article is the typical output of a stationary source: the author, then in principle all we need do is calculate $p(j|i_1,i_2,\ldots,i_N)$ for all letters j and for all N-tuples of letters and blanks which might precede j. From this we then compute

$$F_N = -\sum_i \sum_j p(b_i,j) \log_2 p(j|b_i)$$

where b_i denotes a typical block of N-1 successive letters preceding j. Were these F_N known, then we could estimate the entropy of the sample to any desired accuracy using the fact that

$$H = \lim_{N \to \infty} F_N.$$

The difficulty becomes apparent when it is realized that from a 27 letter alphabet there are 27^N possible N-grams. Of course, many of these are ruled out as impossible in English, but even were we to assume that, say, only one per cent were possible, there would still be 1,968 cases to be examined with N = 3, and 53,144 for N = 4.

Nonetheless, F_N can be computed for very small values of N, and Shannon [91] reports that

F_1 = 4.14 bits/letter

F_2 = 3.56 bits /letter

F_3 = 3.3 bits/letter

His calculations are based on the letter, digram, and trigram frequencies

which have been prepared for coding work (Pratt [78]). Not only is it practically impossible to carry this approach much further, but Shannon suggests that F_3, and all higher F's, may be liable to some error since many of the N-grams in the sample will bridge across two words. It is clear that other approximate techniques are necessary.

Three proposals have been made. The first employs, in one way or another, the built-in knowledge of English statistics in English-speaking people. The second attempts, [illegible] an assumption, to by-pass the sampling difficulties of the direct procedure discussed above. The last utilizes the known empirical distributions of English words, though ignoring the statistical dependencies among words, to determine an upper bound on the entropy. We shall discuss the proposals in this order.

2.1 Shannon's Upper and Lower Bounds

In his original report, Shannon [pp. 25-26, 88] states that "The redundancy of ordinary English, not considering statistical structure over greater distances than about eight letters is roughly 50 per cent." (The definition of redundancy was given in section I.3.6.) In a later paper [91] he cites his original estimate as about 2.3 bits/letter. He arrived at this figure using two techniques. First, he developed approximations to English using the published frequencies, digram, and trigram frequencies of letters and the frequencies and digram frequencies of words to generate approximations to English. The redundancies in each case were calculated; in the last two cases some extrapolation was required, since the tables were not complete.

Second, he selected passages of English at random, and using a table of random numbers he deleted (but with an indication that a deletion had occurred) a certain percentage of the letters. His subjects then attempted to reconstruct the original passage, and he found that the letters could be restored with high accuracy when 50 per cent were deleted, from which he concluded that the redundancy must be at least 50 per cent.

In this second paper [91], Shannon carries his estimation procedures further by developing both upper and lower bounds for the entropy, and his data indicate that the redundancy may be nearer 75 per cent than 50 per cent. He selected 100 samples of English text, each consisting of 15 letters. A subject was required to guess at the first letter of a passage until he obtained it correctly. Knowing it, he guessed at the second until it was obtained. In general, knowing N-1 letters he guessed at the N^{th} until he was correct. The data may be presented as a table having 15 columns and 27 rows (26 letters and a blank). The entry in column N and row S is the number of times subjects guessed the correct letter on the S^{th} guess given that they know the N-1 preceding letters. A small portion of the table is reproduced:

		N					
		1	2	5	10	15	100
	1	18.2	29.2	51	67	60	80
S	2	10.7	14.8	13	10	18	7
	3	8.6	10.0	8	4	5	-

The column marked 100 was obtained by presenting the subject with 99 letters from a 100 word passage. The data for columns 1 and 2 were prepared from published word and digram frequencies which are based on far larger samples.

To use these data, Shannon introduced the notion of an ideal predictor who, knowing $p(b_{i,j})$, i.e., the probability of all N-grams, would select letters j in order of decreasing probability for the given b_i. Thus each letter of a message can be replaced by a number between 1 and 27 which tells how many guesses will be needed before the correct letter is obtained. For an ideal predictor this sequence of numbers will contain the same information as the message, since one can be constructed from the other, but it has the added feature that there will be limited statistical dependencies among the numbers, since the difficulty of one will not generally determine that of the next. Hence, computing the entropy of the number sequence is not difficult, and it can be used to estimate the entropy of the language.

The frequency of the number k in the reduced text will, of course, be given by

$$q_k^N = \sum p(b_{i,j})$$

where the sum is taken over all (N-1)-grams b_i and over those j's such that it results in the k^{th} largest probability for the given b_i.

Shannon then shows that the N^{th} order entropy, F_N, is bounded by

$$\sum_{k=1}^{27} k(q_k^N - q_{k+1}^N) \log_2 k \leq F_N \leq \sum_{k=1}^{27} q_k^N \log_2 q_k^N .$$

Using the data described above, and smoothing them, Shannon calculated upper

and lower bounds for N = 1,2,...,15, 100. Some of the values are:

	N					
	1	2	5	10	15	100
upper bound	4.03	3.42	2.7	2.1	2.1	1.3
lower bound	3.19	2.50	1.7	1.0	1.2	0.6

Upper and Lower Bounds on F_N

When both sets of points are plotted for N = 1,2,...,15, there still remains some sampling error, but smooth curves can be faired through the points reasonably well.

It should be noted that there is a considerable drop in both bounds between N = 15 (at which point the curves are nearly flat) and N = 100. Whether or not this is meaningful is difficult to say, but, as we shall see, none of the other estimates suggests that the entropy is as low as 1.3 bits/letter; however, it must be kept in mind that all of these will be upper bounds, and how much too large they may be is not known.

2.2 The Coefficient of Constraint

Newman and Gerstman [75] approached the problem in another way which does not depend on "built-in" knowledge of English statistics, but which does employ an as yet unproved assumption. They define

$$H(1) = -\Sigma\, p(i)\log_2 p(i)$$

and $$H(1,N) = -\sum_i \sum_j p(i,j) \log_2 p(i,j),$$

where i and j are letters in a passage which are separated by N-1 others. That is, H(1,N) measures the average statistical dependence of a choice j on the choice i which was made N letters earlier. As N becomes large it is clear that this dependence decreases. A measure of its magnitude is

$$H_1(N) = H(1,N) - H(1).$$

They then define a quantity

$$D(N) = 1 - \frac{H_1(N)}{H(1)}$$

which is called the coefficient of constraint. It is a quantity which is 1 when the N^{th} selection is uniquely determined by the first, and 0 when the N^{th} is independent of the first. Since only pairs of letters are involved in these quantities, it is comparatively easy to determine them for a given sample of language.

Using a 10,000 word sample from the Bible, they obtained the following data:

	N					
	2	3	4	5	6	10
D(N)	.223	.103	.064	.039	.027	.012

and a letter frequency entropy of 4.08, which we observe is slightly different from the 4.14 obtained by Shannon. A plot of these data on log-log paper is approximately linear with a slope of -2.0, or, in other words, $D(N) = 1/N^2$, approximately.

The problem now is whether we can estimate F_N from data on $D(N)$. The answer is 'yes,' provided it is true that

$$F_N \leq [1 - D(N)]\, F_{N-1}.$$

This relation is certainly true when $N = 2$ - indeed, the equality holds then - and it is true for any N such that the symbols are independent, for then $D(N) = 0$ and $F_N = F_{N-1}$. They point out, however, that no proof of the assumption has been found, and they add without further elaboration the cryptic comment "... and there are limiting cases in which it is proved not to apply." [p. 120, 75] In any case, if it is assumed, one has

$$F_N \leq F_1 \prod_{i=2}^{N} [1 - D(i)]$$

$$= H(1) \prod_{i=2}^{N} (i - 1) \prod_{i=2}^{N} (i + 1) \prod_{i=2}^{N} \frac{1}{i^2}$$

$$= H(1)\, \frac{(N + 1)}{2N}$$

where we have introduced the empirically grounded assumption that $D(i) = 1/i^2$. In the limit

$$H = \lim_{N \to \infty} F_N = H(1)/2,$$

which gives an upper bound, if the two assumptions are correct, of 2.04

bits/letter. In addition, for N = 1,2,...,15 they compute $H(1)\frac{(N+1)}{2N}$ and they compare these points with those obtained by Shannon as an upper bound. This curve seems to fit the points as well as the faired curve of Shannon.

2.3 Distribution of Words and Letter Entropy

The third, and last, major approach to setting bounds on the letter entropy rests on a computation of word entropies which is based on known frequencies of word use in the language. This entropy, when divided by the average word length, affords an estimate of the letter entropy which is only an upper bound, since the technique, based as it is only on word frequencies, ignores completely the redundancy due to inter-word influences.

Long before information theory, people had determined the frequency of usage of various words, and it was Zipf [102] who observed that if we rank words 1,2,...,r,... in order of decreasing frequency, then the frequency of use of a word is simply proportional to the inverse of its rank. That is, the probability p_r that a randomly selected word is of rank r is given, approximately, by

$$p_r = k/r,$$

where k is a proportionality factor independent of r. There is a certain ambiguity as to just how many ranks there are and certainly if we consider all possible English words the approximate law fails for very high ranks. The value of k is chosen so the 'law,' known as Zipf's Law, holds for the lower ranks, and the size N of the vocabulary is given by the condition

$$\sum_{1}^{N} p_r = k \sum_{1}^{N} \frac{1}{r} = 1.$$

Newman and Gerstman [75], Miller [67], and Shannon [91] have all carried out this computation, but as Newman and Gerstman point out, there are certain discrepancies in the results. Shannon obtains N = 8,727, while Miller, presumably using a definite integral to approximate the series, gets 22,000, and Newman and Gerstman obtain 12,370 by taking into account the discontinuity of the first 100 ranks and approximating the rest of the series by an integral.

Using this distribution, it is then possible to calculate the entropy of the independent word selections according to the distribution, i.e.,

$$H = - \sum_{1}^{N} \frac{k}{r} \log_2 \frac{k}{r} .$$

Shannon obtains 11.82, Miller 10.6, and Newman and Gerstman 9.7 bits/word. These give estimates of 2.62, 2.36, and 2.16 bits/letter if we take 4.5 letters to be the average word length. There appears to be a further disagreement, as was pointed out by Newman and Gerstman [p. 124, 75]. Considering the different values of N obtained, both the Shannon and the Newman and Gerstman results should be on the same side of the Miller result; they are not.

Another approach to the problem from the point of view of words is due to Bell [2]. He supposes that the space between words is sent infallibly and then he observes that the length of a word carries some information. "As

the simplest example, consider the fact that there are only two words of one letter in normal use: the personal pronoun 'I' and the indefinite article 'a.' Hence only two out of the 26 single-letter 'words' which are mathematically available from the alphabet are admitted to the English language, and it follows that when a word of one letter is received in English the choice is only 1 out of 2 instead of 1 out of 26. An alternative expression of this is that the 'internal information' implicit in the fact that the 1-letter word is in the English language equivalent to a selection of 1 out of 13 alternatives; and the communication of a selection of 1 out of 13 would be regarded as a communication of 3.7 'bits' of information ($\log_2 13 = 3.7$), so that the average internal information of 1-letter words in the English language may be stated as 3.7 bits per letter." [p. 384, 2]

For longer words such a detailed analysis is impossible, so he made statistical samples from the dictionary. From this he calculated the internal information in bits/letter and he obtained:

	Number of Letters							
	1	2	3	4	5	6	7	8
Internal Information	3.7	2.2	1.53	1.93	2.36	2.66	2.98	3.21

This curve was smoothly extrapolated for words longer than 8 letters. Using Dewey's word list [11] to obtain relative frequencies of words of various lengths he calculated the weighted average of the internal information and he obtained 2.1 bits/letter.

2.4 The Role of Redundancy

Whatever the correct value of the letter entropy is, it is clear that it is not much over 2 bits/letter and not much less than 1, and so the redundancy is somewhere between 50 and 75 per cent. In other words, we could transmit the same information as we do either by using a considerably smaller alphabet and keeping the length of books and articles the same, or by keeping the same number of symbols in the alphabet and reducing sentences and books to from one quarter to one half their present length. That our language is not fully efficient in this statistical sense presumably results from our need to communicate rapidly and accurately under adverse conditions, i.e., where there is noise: in the presence of other voices, in the wind, at sea, etc. It is clear from the little example given in section I.4.3 that even a small amount of noise can result in a serious drop in the information transmitted - in that case a one per cent chance of error resulted in a ten per cent drop in the entropy. It thus appears reasonable that if a language is designed to cope with even a slight amount of noise, then the redundancy must be quite high indeed. Of course, when the noise level is so high that the natural redundancy of the language is unable to combat it, other methods are used, e.g., words and even whole sentences are repeated, and in such places as factories the vocabulary between two people may be reduced to a few words - possibly, to 'stop' and 'go.'

An example of a purposeful increase in redundancy is found in the very formal language used for air traffic control at an airport. Frick and

Sumby [20] have presented a summary of their findings for this language, but without much of the data. They used the technique, introduced by Shannon [91], of having subjects predict the next letter of a message. Using trained personnel as subjects they found that the uncertainty of control tower language is about 28 per cent that of random sequences of letters and spaces. And this, they point out, is a serious overestimation, since in practice the operator almost always knows the pilot's situation and therefore certain messages are excluded. To estimate these situational constraints, they described hypothetical situations to 100 Air Force pilots and asked them to predict the control tower message. Forming equivalence classes of 'meaning units' and taking into account the imposed grammar of the language, they found that the uncertainty was no more than 20 per cent of what it would have been had the units been equally likely and randomly selected. The overall effect, they estimate, is a redundancy of about 96 per cent. This is not an implausible result when one considers the high noise level in both the tower and the plane, and especially the low margin of allowable error.

A similar study of tower-pilot communications at the Langley Air Force base has been presented by Felton, Fritz, and Grier [18]. As in the Frick and Sumby work, they divide messages into information elements - "... a word or a group of words representing a type of information, such as runway assignment, elapsed time, etc." [p. 5] They divided the analysis of redundancy into three levels: first, they simply took into account the frequencies of the various information elements; second, they determined the predictability within a message; and third, they determined the predictability

between messages from the observed conditional probabilities between messages. At the second level, they determined the probability of each message and determined the entropy of whole messages. This divided by the average number of elements per message was taken to be the entropy of each element. A justification of this procedure was given. The data are separated into messages originated in the air and at the tower, and the estimated redundancy using each of the three levels is presented:

	Level		
	1	2	3
Air	.35	.72	.81
Tower	.26	.75	.78

Redundancy

The authors estimate that if contextual constraints are taken into account, as they were in the Frick and Symby paper, then the redundancy would be about 93 per cent, which compares closely with the 96 per cent mentioned above.

3. Distribution of Words in a Language

In the last section we used the empirically grounded observation of Zipf that if the words of a natural language are ranked from the most to the least common then the frequency of the r^{th} word is approximately inversely

proportional to r. Zipf found that more linguistic data could be fit by the more general equation

$$p_r = P_r^{-B}$$

where p_r is the frequency of the r^{th} word and P and B are constants, B being in the neighborhood of 1 for all languages. "Although this relation appears with regularity in linguistic data, no one has claimed more than a vague appreciation of its cause of significance. No one, that is, until Mandelbrot." [p. 443, 72] Mandelbrot has discussed his work in several places,[57, 58, 59, 60], the clearest probably being [60], and Miller [72] has given a very helpful summary of it.

Mandelbrot started with the assumption that the language - like all known ones - is discrete, i.e., that communication is by means of units called words which are separated by a space. He further assumed that the transmitter in the communication system encodes and the receiver decodes word by word. "Although it may seem trivial, the introduction of the space between words is the crux of Mandelbrot's contribution and the main feature that leads him to results different from Shannon's. In Shannon's problem, the entire message is remembered and then coded in the most efficient form for transmission. In Mandelbrot's problem, the message is remembered only one word at a time, so that every time the space occurs the transmitter makes the most efficient coding he can of that word and then begins anew on the next word. Obviously, a transmitter of the kind Shannon studied will be more efficient, but one of the kind that Mandelbrot is studying will be more practical." [p. 444, 72]

Let us assume that the words are ordered by decreasing frequency of occurrence; denote them by $W_1, W_2, \ldots, W_R$. Let the corresponding frequencies of occurrence be $p_1, p_2, \ldots, p_R$. Let us suppose that to each word there will be a cost C_r for using it - we do not specify what we mean by cost except that it can be summarized by a real number. It might be the number of bits required to transmit it, or the delay, etc. The first problem Mandelbrot attacked, which he called the 'direct problem,' is to find what the costs C_r should be so as to result in the least costly transmission of messages assuming word-by-word coding and known frequencies p_r. This condition yields, as a first approximation,

$$C_r = [\log_M r]$$

where $[x]$ denotes the next integer following x. A better approximation is

$$C_r = [\log_M (r + n) + \log_M d]$$

where M, n, and d are constants independent of r. Observe that the cost depends on the ranking, but not on the details of the probability distribution.

Next, we turn to what Mandelbrot called the 'inverse problem.' In the problem he assumed the words given and their costs fixed, and the task was to determine the frequency distribution p_r such that some economy criterion is met. He has given several criteria which all lead to essentially the same result.

1. Let us suppose that the average cost per word,

$$C = \sum_{r=1}^{R} p_r C_r,$$

is fixed in advance, and we look for the best frequency distribution to transport information (in Shannon's sense). That is, we maximize $H = -\Sigma p_r \log p_r$ subject to the above constraint. (This problem is formally identical to Boltzman's problem in statistical mechanics: to find the maximum entropy for a given average energy.) The following conditions are necessary and sufficient to solve the problem:

$$p_r = P' M^{-BC_r}$$

$$B > 0$$

$$\Sigma p_r = 1$$

$$\Sigma p_r C_r = C$$

The third condition determines P' and the fourth B, provided that $C < \log R$. Note the cost C_0 of the space does not enter here.

2. A second condition, which is a trivial modification of the first, is to hold H fixed and choose the distribution so as to minimize the average cost C. The only difference that results is that B is determined by the value of H, provided $H < \log R$. Again the value of C_0 is irrelevant.

3. A more interesting variant is to allow R and C to be free and to minimize the average cost per unit of information: i.e., minimize

$$\frac{\Sigma p_r C_r + C_0}{-\Sigma p_r \log p_r}$$

subject to the constraint $\Sigma p_r = 1$. As before, we find that

$$p_r = P'M^{-BC_r}$$

but now B is determined by the value of C_0, and so both the value of C and of H are fixed by the choice of C_0.

Finally, we turn to what Mandelbrot called the 'secrecy problem.' He supposed that the words are composed of letters $L_1, L_2, \ldots, L_G$, where G is much smaller than R. Let the letters be labeled in order of decreasing frequency, denote the frequency distribution by q_i, and write the cost of the i^{th} letter as c_i. The cost of a word is assumed to be given by the sum of the costs of its component letters.

"The best possible of all weighted vocabularies from the point of view of the secrecy encoder is the one in which the most economical code is also unbreakable. The code must then be a random sequence of elements, space included, and the enemy must either go to word relationships, that is go beyond our approximation, or try all keys, the number of which is astronomical." [p. 131, 60.] The requirement he places is that an unbreakable random sequence of letters transport information for the smallest possible cost per unit of information. This is similar to condition 3 of the inverse problem, differing however in that there is no element corresponding to the word space. Formally, the condition is that

$$\frac{\Sigma q_i c_i}{-\Sigma q_i \log q_i}$$

should be a minimum subject to the condition that $\Sigma q_i = 1$. From this requirement it can be shown that the word distribution must be

$$p_r = P'M^{-BC_r}$$

as before, but with the added conditons that $B > 1$ and $R = \infty$. The latter condition follows from the requirement of a random sequence of letters to sustain secrecy. We shall discuss the condition $B > 1$ a little later.

Let us summarize: to attain the least costly transmission when words are ranked in order of decreasing frequency, then

$$C_r = [\log_M (r + m) < \log_M d].$$

To attain 1) the maximum information transport with the average cost per word fixed, or 2) the minimum average cost per word with the information transported held fixed, or 3) the minimum average cost per unit of information, then the distribution of the words should be

$$p_r = P'M^{-BC_r}.$$

If we combine these two conditions, taking into account the fact that statistical fluctuations in data will smooth over the steps of the former equation, we obtain

$$p_r = P(r + m)^{-B},$$

which Mandelbrot has called the 'canonical curve.' Observe that if $m = 0$, this is the generalized Zipf law.

As Mandelbrot points out, the fit of Zipf's law to most language data is good only in the central range and it is in error for the most frequent and the least frequent words. By choosing values of B and m different from 1 and 0 he has been able to achieve far better fits.

The condition $B > 1$ which results from the secrecy criterion has been found to be met by most natural languages. Zipf called those with $B > 1$ 'open vocabularies' and those with $B < 1$ 'closed vocabularies.' Most languages with closed vocabularies are in some way peculiar or special.

Clearly, Mandelbrot's theory, like Shannon's, is normative, but it is much more closely related to a specific empirical field than is Shannon's. Thus the question must be raised as to exactly what Mandelbrot has shown and what it means for linguistics. "He says that if one wants to communicate efficiently word-by-word, then one must obey Zipf's law. There is a strong temptation to reverse the implication and to argue that because we obey Zipf's law we must therefore be communicating word-by-word with maximal efficiency." [p. 445, 72] Of course, Miller goes on to point out that much other evidence exists - such as the redundancy data discussed in the last section - to suggest that this reversed implication is false. It remains to be seen whether it can be shown that marked deviations in certain directions from perfect efficiency result in only slight deviations from the canonical curve.

4. The Capacity of the Human Being and Rates of Information Transfer

In recent years it has proved necessary to construct a variety of complex information-processing systems in order to deal with certain

military and industrial problems. These systems typically receive, from diverse sources, a tremendous amount of raw information which must be filtered, recoded, and correlated into what may be called a model of some situation of interest. The model must be sufficiently simple so that a person can grasp it completely, and sufficiently accurate to that he can reach useful decisions on the basis of it. For example, an air defense system receives raw information from radars, spotters, airline schedules, weather reports, fighter readiness reports, etc. All of this must be reduced to a simplified model of the enemy attack, the defense facilities, and the defensive response, so that a commanding officer, with only a few seconds' or minutes' delay, can know the situation continuously. The officer must make and modify his defensive decisions on the basis of such a model. It is clear that much of this processing - especially where speed and accuracy are needed - can and should be reduced to machine operations, but, with our present technology, there are certain steps which are far more simply and effectively carried out by a person than by a machine. For example, one of the first steps in an air defense system, and one which is not easily duplicated by a machine, is the isolation and transfer of pertinent information from a radar scope face. From all the random noise and background reflections on the scope an operator must single out those 'blips' which are aircraft, and this he must introduce into the rest of the system, say, as a [illegible] telephone message. The question arises as to how much information he can process per second over a sustained period.

It is clear that for any specific problem of this type, an answer

can be obtained by direct experiments on the trained personnel using the equipment. On the other hand, the question arises whether it is necessary to study each new situation separately, or whether the pertinent variable is the amount of information in bits/sec which will be presented to the operator as compared with the maximum amount he can handle.

That is, can we treat a human being as a channel and so determine a channel capacity for him? If this is possible, it will certainly simplify the design problem, for it is generally not too difficult to determine the rate of information flow in the machine components of a system. The question of whether it is useful to treat men as channels in certain situations remains, in the opinion of many, still an open problem. This is not our question here; we need only recount some of the studies which have been executed to determine his capacity under the assumption that he can in fact be usefully considered as a channel.

Considering the theory presented in part I, two procedures to estimate the capacity seem possible. First, from whatever physical, physiological, and psychological facts known and relevant to the type of transmission being employed, to make an estimate of the channel capacity. Second, by varying certain variables and by employing diverse coding schemes, to find the maximum amount of information which he can be caused to handle. This, by the fundamental theorem of information theory, affords a lower bound on the capacity. Roughly speaking, the first procedure has resulted in upper bounds of the order of 10,000 bits/sec, while the second yields a lower bound somewhere in the range of 10 to 100 bits/sec. The consensus is that the

lower bound more nearly represents the human capacity, but no really strong argument exists to support this view except that no one has yet devised a way to achieve a higher rate. We shall now examine these estimates in a little more detail.

4.1 Upper Bounds

Possibly part of the difficulty in obtaining a satisfactory estimate using the first procedure is the present lack of an adequate model for what happens functionally within a person when he is processing information. Thus, independent measurements on most of the 'channel' - which is surely not homogeneous in its properties - cannot be had. As a result, the estimates which have been made are in a sense only concerned with the peripheral aspects of the channel. We will cite in a moment another reason which has been offered to explain the difference between the upper and lower bounds.

Licklider and Miller [53] have pointed out that an estimate of the capacity with respect to auditory signals can be obtained from a result of the theory of information for continuous systems (see the appendix). It is known that if the bandwidth of the channel is W cycles/sec, and if the noise and the signal are simply additive with a power ratio of P/N, then the capacity in bits/sec is given by

$$C = W \log_2 \left(1 + \frac{P}{N}\right).$$

For auditory signals a bandwidth of 5,000 cycles/sec is conservative and a signal-to-noise ratio of 30 db, or a power ratio of about 1,000, is not unusual, in which case the capacity must be about 50,000 bits/sec. In

actual attempts to transmit selective information by auditory means, a rate as high as 50 bits/sec is unusual. In other words, the efficiency of the auditory system must be considered to be about 0.1 per cent. Licklider and Miller offer the explanation that most of the information transmitted by an auditory signal is personal information about the originator - his way of speaking, his mood, and some of his linguistic history. While this may well be the case, it is interesting that no one has yet devised a way to use this apparently available capacity for the transmission of preassigned selective information.

A far more detailed estimate of auditory capacity has been made by Jacobson [44, 45] using various data about hearing, such as the total number of monaurally distinguishable tones. He concludes from his analysis that one ear should be able to handle about 8,000 bits/sec, and with very loud sounds, 10,000 bits/sec. It is known that there are approximately 29,000 ganglion cells from the ear, hence the average rate of information transfer over a nerve fiber is about 0.3 bits/sec. However, he points out that "It is very unlikely that there is any binary or similar coding in the cochlear nerves. It is consequently not particularly meaningful to state that the average informational capacity of a single cochlear fiber is about 0.3 bits/sec." [pp. 470-471, 45] This result, however, can be translated into the equivalent number of tones which can be distinguished on one fiber, and he obtains 40 tones/sec.

Jacobson [46] has also carried out a similar calculation for the eye, taking into account facts known about discriminability, etc., but

ignoring the effect of color. He obtains an estimate of 4.3×10^6 bits/sec for each eye. From this one can conclude the maximum average rate over each neural fiber must be 5 bits/sec. The inclusion of color would, of course, raise this estimate.

So far as we have determined, these are the only estimates of channel capacity which are based on measurements independent of the actual rate of information flow. We turn now to estimates of how rapidly information of a particular type can be, or rather, has been, caused to pass through a person.

4.2 Lower Bounds: Maximum Observed Rates of Information Transfer

Let us first consider the transmission of language encoded information. Miller [67] points out that if we consider the average measured length of vowels and consonants = about 12.5 sounds/sec -- and if we were to suppose that they are equi-probable and independently selected, then speech would convey information at a rate of 67 bits/sec. If, however, we take into account their relative frequencies (Dewey [11]), then the rate is reduced to about 50 bits/sec. Further, if we take into account the fact that vowels and consonants tend to alternate in English, the estimate is only 46 bits/sec. Finally, on the basis of Zipf's law, Miller estimated that there are 10.6 bits/word (section II.2.3). Since a speaker can sustain a maximum of about 3 words/sec, the transmission rate using speech can be no more than 32 bits/sec. "The maximum efficiency within the restriction imposed by the phonetic structure of English words, therefore, is about 50 per cent." [p. 798, 67]

In practice, however, an ordinary speaking vocabulary is not as large as assumed when Zipf's law is called for, nor can a person usefully employ a speaking rate of 3 words/sec. An assumption of an equi-probable distribution over a vocabulary of 5,000 words which are spoken at a rate of 1.5 words/sec yields an information rate of 18 bits/sec.

In addition, as Quastler and Wulff [82] point out, the various rate estimates using Zipf's law ignore the constraints among words. They cite evidence which suggests that the guessing of a missing word within context may be correct as much as 30 per cent of the time. This reduces the information transmission rate to about 7 or 8 bits/word, and if we assume that 15 per cent of the words are incorrectly received, the estimate must be reduced to 6 or 7 bits/word. Using Miller's speaking rate of 1.5 words/sec, it appears that from 10 to 20 bits/sec is a good average rate of transmission, and that with rapid speech the rate may get as high as 25 bits/sec.

Quastler and Wulff report data on several other methods of information transfer, and in summary they find that 25 bits/sec seems to be the maximum rate. In all cases, a mechanical response was required of the subject, but they verified that mechanical limitations were not determining an apparent rate by showing that higher rates could be achieved if memorized materials were used. The first experiment they discussed was based on typing, but it was known *a priori* that this would not [illegible] fastest possible rates, since text can be read aloud faster than a typist can take it down. For this experiment, random sequences of letters were drawn from

alphabets of 4,8, 16, and 32 symbols. Three typists with from 5 to 12 years' experience were paced by a metronome at 2, 3, 4, and 6 beats/sec. In general, the errors which occurred were the transposition of letters, and so it is a question as to whether these should be treated as one or two errors. Depending on this, we obtain the following upper and lower bounds on information transmitted (section I.5)

	Alphabet size			
	4	8	16	32
Upper Bound	6.7	10.5	13.2	16.7
Lower Bound	3.8	7.4	11.8	13.4

Information Transmitted in bits/sec

It was found, as would be expected, that with the higher metronome speeds and with the larger alphabets, the greater percentage of errors occurred. For 8 and 16 symbol alphabets a speed of 3.2 $\pm$ 0.2 keys/sec represented the highest effective speed, and beyond that their precision so decreased as to keep the transmission rate about constant, and beyond 4.5 keys/sec the quality of their output decreased very rapidly. With 4 symbols the effective speed was 3.6 keys/sec, and with 32 it was 2.9 keys/sec. When the subjects were not driven by a metronome, but were instructed to type as rapidly as possible, it was found that the rate of transmission was down about 9 per cent.

A second experiment drew on the sight-reading ability of three young pianists. They were presented with random music (notes selected using random numbers) and they were paced by a metronome which was gradually increased in tempo over trials. Tape recordings were made and each of the subjects scored each of the tapes for errors. The agreement was fair, but both a low count (errors detected by each subject) and a high count (those detected by at least one) were determined. The information transmitted was computed from the error count and from assumptions about the error pattern. Again, several different 'alphabets' were employed: 3, 4, 5, 9, 15, 25, and 37 keys.

The data show that the highest speed for which the error rate remained low decreases from 7 keys/sec for an alphabet of 3 or 4 keys to 4.3 keys/sec for the 37 key alphabet. This decreased speed, coupled with an increase in error rate, keeps the information transmission rate at about 22 bits/sec over a fairly wide range of speed and alphabet size.

In contrast to the typing experiment, individual differences became apparent when the subjects attempted to exceed their limits. One kept the error rate low by failing to keep up with the metronome, another kept the pace but allowed the error rate to become large, and the third held the pace for periods and then he would lose the beat.

A third set of materials for determining capacity which Quastler and Wulff have studied is mental arithmetic problems. They point out that if certain plausible assumptions are made about the information involved in calculations, and if the published time data on so-called 'lightning

calculators' (people who are noted for rapid mental calculations) are used, one obtains an estimate of 22 to 24 bits/sec for the transmission rate. The feat of such people appears, therefore, not to be a high rate of information transmission, but rather a tremendous storage of information for short periods of time. In addition, Quastler and Wulff conducted some simple experiments on mental addition of columns of figures. On the average they found - again by making some plausible, but debatable, assumptions - a rate of 6 to 12 bits/sec, but one exceptional subject sustained a rate of 23 bits/sec.

From these data, and others not published, it appears that it is difficult to cause a subject who is employing familiar operations to exceed - let us be generous - 50 bits/sec, even though present estimates of ear and eye capacity exceed this several hundred times. It certainly seems an open problem to bring these two estimates closer together, either by devising a method to employ much more of the apparent capacity to transmit selective information, or by a more detailed analysis of the human being as a channel to show that 50 or 100 bits/sec is truly his limit. Jacobson's comments on this disparity are of interest. "Thus it is evident that the brain can digest generally less than 1 per cent of the information our ears will pass. It must be appreciated that the ear is a channel vastly wider than its apprehensible output. It is the ability of the brain to scan for those portions of the auditory signal which are of interest which makes the wide capacity of the ear maximally useful." [p. 471, 15]

4.3 Other Observed Rates of Information Transfer

Not all the experiments, or the observations taken, on rate of information transfer have resulted in rates as high as those described a ve. Evidently the mode of presentation of the information vitally affects the rate at which it can be handled; if this conclusion is true, then the naive program outlined at the beginning of this section must be modified to some dgree.

In this connection the results of an experiment performed by Klemmer and Muller [49] are of interest. The stimuli consisted of five lights arranged in an arc; a corresponding set of telegraph keys was arranged under the subject's fingers. The subject was to press the keys corresponding to those lights which were on. By using various numbers of bulbs - the subjects were told which would be employed - 1, 2, 3, 4, and 5 bits could be achieved in the presentation. In addition, the stimulus cycle, which consisted of lights on 50 per cent of the cycle and off the last 50 per cent, was presented at a rate of 2, 3, 4, and 5 cycles per second. The subjects were all trained on the apparatus for several weeks, and the practice curves indicate that they had completely stabilized by the time the experiment was performed.

For a fixed number of bits in the stimuli, it is found that by varying the rate of information presented there is a nearly linear increase in the transmitted information until a peak is reached, after which the transmission rate falls markedly. The location of the peak, and hence its value, is an increasing function of the number of bits in the stimulus.

The approximate values of the peaks are:

	Information presented in bits/stimulus				
	1	2	3	4	5
Peak Transmitted Info. in bits/sec	2.7	4.0	5.8	8.4	10.5

The decay of the performance following the peak is remarkable. In the case of a stimulus with 5 bits, the peak of 10.5 bits/sec occurs when the input rate is approximately 13 bits/sec. When the rate is increased to 15 bits/sec, the transmitted information has dropped to 6 bits/sec. This drop is, of course, due to a radical increase in the error rate.

It should be mentioned that what we report are average results, and the authors present data to show that there is considerable individual variation.

Now, it is clear that the maximum rates found in this experiment are less than those described in section II.4.2 above. In many respects this experiment and its conclusions are more closely related to those described in the next section on reaction times than it is to either the reading, typing, or music experiments. One important difference is that in the latter experiments the stimuli are before the subjects at all times and hence the receptor mechanism can operate with a considerable lead over the response mechanism, whereas such a large lead was not possible in Klemmer and Muller's study. It therefore appears to be more nearly a 'continuously' executed reaction-time experiment. This can be supported from data they present.

Typical reaction-time experiments were run on the same subjects, and a comparison of the inverse of the reaction time to the stimulus rate (in stimuli/sec) at peak transmission is revealing:

	Bits in Stimulus				
	1	2	3	4	5
1/RT	3.8	2.6	2.6	2.4	2.4
Stimulus rate at peak transmission	3.7	2.4	2.4	2.4	2.4

The Felton, Fritz, and Grier [18] study of communications at Langley, discussed in II.2.4, yields some data on operational rates of information handling. Using 'information elements' on which to base their calculations, they found that during a single landing the following amounts and rates of information were employed by pilots and tower:

	Transmitted in bits	Rate in bits/sec
Air	114	8.4
Tower	133	10.3

However, it will be recalled that they determined that there was a very high redundancy in the transmission, and if only 'new' information is considered the table becomes:

	New Information Transmitted in bits	Rate of new information transmitted in bits/sec
Air	22	1.6
Tower	29	2.2

Either set of rates is below that which we have seen is possible for speech.

Hick writes, "As a personal speculation from such data as are available, it seems likely that transmission rates fall into three fairly distinct classes:-

1. High rates of 10-15 bits per second.
2. Moderate = 5-6 bits per second.
3. Slow - 3-4 bits per second." [p. 63, 35]

He feels that these rates are closely correlated to the mode of presentation of the information. High rates are obtained only through simple 'imitation' codes of the type we learn in childhood. Moderate rates are typical of 'arbitrary' specially learned codes in which each signal has a high information content. The low rates result from arbitrary codes having a low information content per signal and a high rate of presentation. As a partial and speculative explanation for rates less than full capacity Hick comments: "But for various reasons I am inclined to suspect -- I would certainly not be more definite than that - that there is a tendency, overcome, if at all, only with long practice, to sidetrack one or two bits per discrete movement as a kind of monitoring feedback. It would be originally necessary in the course of developing the skill (the code being, as stated above, relatively

arbitrary or 'unnatural'), and may be retained, perhaps as a habit, or perhaps to keep the skill up to full efficiency, for a long time after that." [pp. 70-71, 35]

5. Reaction Time and Information Transfer

Our present topic may, in a sense, be considered a continuation of the last section on capacity; here we shall deal with what might be called 'momentary' capacity. Previously we considered long samples of sequential stimuli to which the subject responded more or less continuously; now we shall consider his reaction time to a single isolated display. The question is what characteristics of the display need be considered in order to account (simply) for the observed reaction times. The hypothesis, very generally, is that the information content of the display is the relevant variable and that the reaction time will turn out to be a very simple function of it - namely, linear.

There are, according to information theory, a number of ways in which the information transmitted can be varied: a) by varying the number of equi-probable alternatives, b) by altering the probabilities of the various choices, c) by introducing sequential dependencies between choices, and d) by allowing errors (noise) to occur. In the theory these are equivalent; whether they produced equivalent human responses is an empirical problem.

In the first experiment of the series of three we shall discuss, Hick [34] considered cases a and d. He presented subjects with a stimulus

in which one of n equally likely alternatives would arise, and the subject had to respond as to which occurred. His hypothesis was that the reaction time (RT) would be proportional to the information in the stimulus, or, in other words, the rate of information transfer would be constant. There is, of course, a difficulty in assuming RT = k log n, since when n = 1 this would require a zero reaction time. Hick suggests that there are really n + 1 alternatives, since we have ignored the case of no stimulus. While this seems reasonable, it is difficult to accept his assumption that all n + 1 are equi-probable and that RT = k log(n + 1). However, he finds that data taken by Merkel [64] are well fit by choosing k = 0.626 and that his own are fit with k = 0.518. Since a fixed delay, independent of n, seems plausible, the function c + k log n might seem intuitively more suited to fitting the data, but it does not fit either set of data as well. These fits were obtained with n in the range 1 to 10, i.e., up to a little more than 3 bits.

Turning to method d of varying the information, Hick points out, "... if the subject can be persuaded to react more quickly, at the cost of a proportion of mistakes, there will be a residual entropy which should vary directly with the reduction in the average reaction time." [p. 15, 34] An experiment was performed in which the subjects were pressed, and the errors were taken into account by computing an equivalent error-free n, n_e. The reaction time data when plotted against n_e were found to be fit pretty well by the curve obtained for the errorless case.

As Hick's student Crossman states, "The original evidence that

the information measure was the appropriate one to use for interpreting choice-reaction times was simply that the logarithmic function occurs in both. This in itself is not strong, since logarithmic relations occur rather often in biological measurement. The case became much stronger with Hick's finding that the reduction in response-time where errors are permitted obeyed the same law." [p. 41, 10]

In Hick's experiment the rate of information transfer was about 5.6 bits/sec, a value which is low compared with the largest obtained using a 'continuous' stimuli presentation.

Hyman [41] has examined methods a, b, and c of varying the information when the performance was kept errorless. He states his hypotheses as

"1) Reaction time is a monotonically increasing function of the amount of information in the stimulus series.

"2) The regression of reaction time upon amount of information is the same whether the amount of information per stimulus is varied by altering the number of equally probable alternatives, altering the relative frequency of occurrence of particular alternatives, or altering the sequential dependencies among occurrences of successive stimuli." [p. 189, 41]

The stimulus presentation was by means of a matrix of lights with a range of 0 to 3 bits. The subjects responded by means of a vocal key, which seems to yield more precise measurements than the hand-operated key of Hick's experiment. The subjects were given complete statistical information about the stimulus and before each test run they were given sample

sequences formed according to the appropriate statistics. Four subjects were used. The correlations reported below are the average of the four correlations computed for each subject separately.

In the first phase, the number of equi-probable alternatives were varied and a correlation of 0.983 was found between reaction times and information in the stimuli. This confirms Hick's results. In the second phase, when the relative frequencies were changed, an average correlation of 0.975 was found. In the third phase, introducing sequential dependencies resulted in a correlation of 0.938. The last correlation is significantly lower than the other two.

Hyman concludes from his data that his second hypothesis, while not acceptable at the 1 per cent level, is acceptable at the 5 per cent level.

In discussing the second phase, he points out that the reaction times of the less probable events were much longer than those of the more probable ones, and that the reaction time used is actually a weighted mean of these. Crossman [10] examined this phenomenon in greater detail as another test of the central hypothesis. "When a subject responds to a sequence of signals all of which belong to a known set but some of which occur more frequently than others, his average response-time will be proportional to the average information per signal. This follows from the hypothesis that the subject deals with information at a constant rate." [p. 41, 10] To test this he used a sorting task on ordinary playing cards and by varying the dimensions on which they were to be sorted he was able to examine the reaction times over a range of 0 to 2 bits/card. The correlation

between reaction time and information in a card was 0.66, and when the data are plotted it appears that no simple curve will fit them better than a straight line.

Crossman adduced evidence to show that the deviations from linearity were due to differential difficulties in discriminating the cards in different classes. On the basis of this he made the important observation that there is "... a major difficulty in the use of information theory in psychology, for information theory in the discrete case stated by Shannon says nothing about actual signals and the process of distinguishing them one from another; it deals only with abstract symbols already identified and distinct." [p. 49, 10] This, of course, suggests carrying out a similar experiment using only one dimension of discrimination and causing the entropy to vary along it. This was done and the fit was improved.

On the basis of his data, Crossman concluded "... our hypothesis that rate is constant under variation of relative probabilities is upheld by these observations, with the proviso that 'discriminability' of signals should be equal in a sense yet to be precisely defined." [p. 50, 10]

From these data it seems reasonable to conclude tentatively that the rate of information transfer in a reaction time experiment is constant when the information in the stimulus is in the range 0 to 3 bits. Since this conclusion is not in conformity with the observations made with a 'continuous' stimuli presentation, it would certainly be interesting to see whether the rate remains constant when there are more than 3 bits in the stimulus,

and also to see whether an experiment can be found with the rate constant, but much larger than 5 bits/sec, for the range 0 to 3 bits.

6. Visual Threshold and Word Frequencies

In the last three years there has been a series of experiments relating the visual threshold of word recognition (as given by tachistoscopic measurements) to the frequency of their occurrence. Originally, the program stemmed from work on the Bruner-Postman hypothesis that sentences which relate to things liked are recognized with less difficulty than those relating to things disliked. Evidence has accumulated that the major relation is actually between recognition speed and the frequency of occurrence of the word in the language. Howes [39] cites data involving sentences, and Howes and Solomon [40] similar data involving only words. In the latter case, word frequency counts were obtained from Thorndike and Lorge [97] and there was found to be a correlation of about -0.7 between recognition time and the logarithm of word frequency. Howes [39] and Miller [68] describe data taken by Solomon in which seven-letter Turkish words were used. These were written on cards which the subjects studied. Some words appeared on many cards, others on only a few, so there was differential exposure to these new words. A correlation of -0.96 was found between recognition time and log frequency. King-Ellison and Jenkins [47] repeated Solomon's experiments with some slight variations, including the use of artificial five-letter words, and they obtained a correlation of -0.99. They point out that a relationship to information theory is suggested, namely,

that recognition time is a linear function of the information transmitted by a word. The earlier comment we quoted from Crossman, namely, that logarithmic relations are so common in biology and psychology that more must be established before an information theoretic model is assumed, is relevant here. Further studies appear to be needed.

7. The Information Transmitted in Absolute Judgments

When a subject is required to place stimuli which vary along one dimension, such as size or loudness, into N simply ordered categories, such as the first N integers, then he is said to be making absolute judgments of the dimension of the stimuli. For example, the stimuli might be pure tones at 100, 150, 200,...,1,000 cycles/sec. Each time a tone is presented he must place it in a category as accurately as he can. It is clear that in general errors will occur of the form: a tone with a lower frequency than another will be put in a higher number category. It is also clear that the error rate can probably be diminished by reducing the number of categories. For example, if he must place the above stimuli in 21 categories, we may expect more errors than if he need only report whether a signal is below or above 500 cycles/sec, for then there will be little ambiguity in his mind except for those stimuli near 500 cycles. Such experiments have a long history, but there has always been some difficulty in summarizing the data - just how should the error picture be summarized?

Garner and Hake [27] pointed out that the matrix relating input stimuli to response categories, with the entries the frequencies of pairings

between stimulus and a category, can be treated (with the obvious normalization) as a noise matrix for a communication system, where the communication is of selective information from the stimuli to the experimenter via the subject as a channel. We may, therefore, compute the information of the stimuli set (which, of course, depends on the relative frequencies of presentation of the different stimuli) and the equivocation of the transmission, and the difference is the information transmitted. If for a certain type of absolute judgment it is found that 20 categories allow the transmission of 3 bits, then in principle as much can be transmitted using only 8 unambiguous categories. Choosing the categories so that there is no ambiguity, i.e., no errors, may be difficult, but Garner and Hake point out that if the errors have a Gaussian distribution the condition is equivalent to a criterion of equal discriminability.

In another paper [30] they cite the major difference between the usual error analysis for experiments of absolute judgments and the proposed information theory analysis. An error analysis ignores the fact that if the error distributions do not overlap, there will be no ambiguity. The information analysis takes this into account, but, unlike the error analysis, it completely ignores the magnitude of the errors. There are some applications where it is preferable to have a multitude of small errors, provided that there is never a single major one.

A number of applications of this proposal have been made to different classes of absolute judgments. Pollack [76] studied tones which were spaced equi-distantly on a logarithmic frequency scale from 100 to 8,000 cycles/sec.

The subjects had to assign a number to each tone presented. When there were 2 and 4 tones in the stimulus set, the transmission was perfect, 1 and 2 bits respectively. But with 8 and 16 tones, the curve became flat, and the average maximum transmission was 2.3 bits, or the equivalant of perfect identification among 5 tones. The best subjects reached the equivalent of only 7 tones. On the grounds that there are known to be 40 to 60 identifiable sounds associated with speech and music, Pollack felt that there must have been a serious underestimation of the information transmitted, and so he performed a series of auxiliary experiments to attempt to raise the value. Six different partitions of the frequency space were examined, and the frequency range was varied with the bottom held at 100 cycles/sec and the top moved from 500, 2,000, 4,000, and 8,000 cycles/sec. These variations resulted in only a few percentage points change in the information transmitted. He suggests that the result is so low because of the acute sensitivity of the information measure to error, which we have mentioned earlier (section I.4.3).

Halsey and Chapanis [32] have presented similar data on the number of absolutely identifiable spectral hues, and though they do not apply an informational analysis, their findings are of some interest. The colors were identified sequentially from violet to red by numbers, and the subjects were familiarized with the number-color code until learning was completed. In a test using 10 hues and 20 judgments per hue, they found that two observers were correct in 97.5 per cent of the judgments. These hues were selected on the basis of several earlier experimental runs in which more hues were employed, but a lower accuracy was obtained. They note that

absolute identifiability of 10 hues is considerably better than had been previously reported, but they attribute this mainly to different experimental conditions.

Hake and Garner [30] applied the information theory analysis "... to determine the minimum number of different pointer positions which can be presented in a standard interpolation interval to transmit the maximum amount of information, not about which positions of the pointer are occurring, but about the event continuum being represented." [p. 358, 30] Two variations were run: in the limited response case the subjects were told the values the pointer could assume and they were required to respond only with those numbers; in the unlimited response case no such restriction was made. 5, 10, 20, and 50 possible pointer positions were used, and the data are summarized below:

		Number of Positions			
		5	10	20	50
Information Transmitted in Bits	Limited Response	2.31	3.14	3.16	3.19
	Unlimited Response	2.29	3.03	3.11	3.11

We observe that beyond 10 pointer positions the amount of information transmitted is roughly constant - equivalent to about 10 errorless positions. There seems to be little or no difference between limited and unlimited responses as far as this analysis is concerned, but Hake and Garner point out that an error analysis shows that the errors increase when the subjects

are allowed unlimited response.

In a later paper, Garner [28] comments: "A measure of information transmission provides a means of specifying perceptual and judgmental accuracy in situations where absolute judgments about various categories on a stimulus continuum are required. This measurement allows the determination of the maximum number of stimulus categories which could be used with perfect accuracy without the necessity of sampling all the possible numbers of categories. However, this use of information transmission requires the assumption that the inherent judgmental accuracy is independent of the number of stimulus categories used experimentally. Two experiments (Garner and Hake and Hake and Garner) have shown that this assumption is quite valid for situations involving judgments of position in visual space, and Pollack's experiment demonstrates its validity for judgments of pitch." [p. 373, 28] Garner then proceeded to examine its validity in judgments of loudness of tones using 4, 5, 6, 7, 10, and 20 categories. He found that judgment accuracy was nearly perfect for 4 and 5 categories (perfect being 2 and 2.32 bits respectively), but that it had dropped to 1.62 bits for 20 categories, which is equivalent to perfect accuracy for only three categories. Thus the assumption is apparently not valid for loudness.

He went on to show, however, that the information transmitted could be improved if both the observers, i.e., the subjects, and the stimuli were taken as inputs to the system and the responses as outputs. (See section I.5.2 for the analysis procedure when there are more than two dimensions.) In other words, there was considerable variability among the

subjects when a large number of categories was employed. A further raising of the information transmitted, so there is no drop at all, is achieved if the stimuli, the observers, and the preceding stimulus are all taken as inputs to the system.

Klemmer and Frick [48] carried out a similar experiment and analysis but with two and three stimulus dimensions instead of one. They flashed (0.03 sec) a display consisting of white dots on a black background to subjects who marked on answer sheet grids what they thought the position of the dots to be. The experiment was run both with and without grid lines on the black background, and there was not found to be an appreciable difference in the data. With the situation restricted to the presentation of one dot, the information in the stimulus could be varied by changing the order of the matrix of possible positions. From 3.2 bits (order 3) to 5.2 bits (order 6) there was an increase in information transmitted from 3.2 to 4.4 bits. From 5.2 bits to 8.6 bits (order 20) in the display, the information transmitted remained approximately constant.

In addition, the number of dots presented was varied, and it was found that by using 4 dots and a matrix of order 3 (7.0 bits) 6.6 bits were transmitted. Further, when from 1 to 4 dots were used, then a display having 8.0 bits resulted in almost perfect transmission - 7.8 bits. "It is clear that the maximum amount of information that can be assimilated from a brief visual exposure is a function of the type of encoding used. The question immediately arises as to whether or not there is a common metric which may be applied to the different message classes and which will correlate

with the maximum information-carrying capacity of that class." [p. 18, 48] They observe that using only one dimension or coordinate (the location of a point on a line) Hake and Garner found a maximum transmission of 3.1 bits, using the two coordinates of a matrix they obtained a maximum of 4.4 bits, and using the two coordinates of a matrix plus the one of the number of dots, they found 7.6 bits transmitted. This suggests that the maximum increases with the number of dimensions.

In this connection, Christie and Luce [9] have suggested that a careful analysis of the distribution of disjunctive reaction times in simple choice situations - like the ones described above - may permit a model of the 'mental' or 'internal' structuring of simple decision processes. They suggest representing this structuring by a flow diagram (also called a network or an oriented graph) which indicates the general temporal organization of certain gross internal processing of the information. Two special and extreme cases are serial and parallel processing, which are diagrammed in the figure. In some highly speculative comments they suggest that parallel processing may be carried out on information which is presented in what we intuitively call several different dimensions, and that serial processing is effected on information lying in one dimension. With some simple assumptions, they show that such a model has the appropriate information transmission properties for a matrix display -

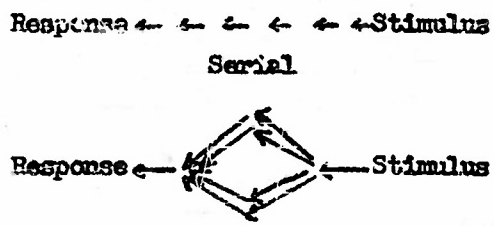

Fig. 5

at least up to 6 or 8 bits. If the techniques they suggest - which we shall not detail here - is practical, then it may serve to give an empirical definition of what psychological dimension means.

On the basis of the several experiments we have discussed, one can conclude that for objective ratings there is, up to a point, an increase in the information transmitted with an increase in the number of categories, and after that point the information transmitted either remains constant or decreases. Bendig and Hughes [3] raised the question whether the same conclusion is possible for ratings of subjective feelings. To study this, they had subjects evaluate, according to either 3, 5, 7, 9, or 11 categories, their knowledge of 12 different countries. Anchoring statements of the form "I know (a great deal) (something) (very little) about this country" were employed in three variations: center anchored, both ends anchored, and both ends and the center anchored. Information transmission, they found, was increased by an increase in the verbal structuring of the scale, i.e., by the anchoring, but the increase was not very marked. With the anchoring held constant, there was a nearly rectilinear increase of information transmitted with an increase of number of scale categories, except that there was a deceleration in the step from 9 to 11 categories. This effect is in accord with the diminishing return observed for objective scaling.

8. Sequential Dependencies and Immediate Recall, Operant Conditioning, Intelligibility, and Perception

One of the main points of the 1949 Miller and Frick [65] paper was to bring to the attention of psychologists that in information theory they

had a tool ideally suited to the characterization of sequential dependencies in the stimulus, in the response data, or in both. There appear to have been four areas of psychological study to which this observation has been applied: to the learning of written material as a function of the statistical dependencies in those materials, to the sequential responses obtained in operant conditioning, to the intelligibility of verbal material as a function of statistical dependencies within the material, and to the ability of subjects to perceive statistical dependencies in materials. We shall discuss them in that order.

6.1 Immediate Recall

"Briefly stated, the problem... is, How well can people remember sequences of symbols that have various degrees of contextual constraint in their composition? The experimental literature contains considerable evidence to support the reasonable belief that nonsense is harder to remember than sense. This evidence has suffered, however, from a necessarily subjective interpretation of what was sensible." [p. 179, 66] Using Shannon's method, Miller and Selfridge [66] prepared N^{th} order approximations to English in the following manner. A sequence of N successive words was chosen at random from a connected text, and a subject was asked to imbed the passage in a meaningful sentence. The first word in his sentence following the original group of N words was recorded. To the next subject was presented the last N-1 words of the original passage plus the new word, and he placed this N-word passage in a sentence. The first word after the passage was recorded, and so on. In this manner they generated approximations

of order 0,1,2,3,4,5, and 7 in passages of 10, 20, 30, and 50 words in length. Using these approximations to English, plus meaningful text, a standard recall experiment was executed. With the passage length held constant, they found that the percentage of recall increases with an increase in the order of approximation to English. In particular, for the 30 and 50 word passages the recall of the 5^{th} and 7^{th} order approximations to English is very little different from the recall of text material of the same length - this notwithstanding the fact that the 5^{th} order is quite nonsensical and the 7^{th} order would by no means be considered English. With shorter passages, recall comparable to that of text was achieved for even lower values of N.

"The results indicate that meaningful material is easy to learn, not because it is meaningful per se, but because it preserves the short range associations that are familiar to the Ss. Nonsense materials that retain these short range associations are also easy to learn. By shifting the problem from 'meaning' to 'degree of contextual constraint' the whole area is reopened to experimental investigations." [p. 183, 66] For example, one may ask whether their conclusion is valid for the whole memory decay curve, or whether it holds only for short term memory.

Similar results have been found by Aborn and Rubenstein [1] in a slightly different experimental situation. They devised an 'alphabet' of 16 nonsense syllables which fell into four easily distinguished classes of four syllables each; this classification was shown to the subjects. From these syllables six classes of passages of 30-32 syllables were constructed. The members of the first class were formed by random selection of syllables,

and the others had increasing amounts of organization. For example, class four passages were marked by commas into groups of four syllables, and the first syllable of each group was chosen from class one, the second from class two, etc. The subjects were allowed 10 minutes to study the formal organization of the passage on which they would be tested and then three minutes to learn the actual passage, after which they were asked to reproduce it as accurately as possible. The authors had two hypotheses: "(a) The amount of learning in terms of syllables recalled is greater as the organization of the passage is greater, i.e., as the average rate of information is smaller. (b) The amount of learning in terms of the information score, computed as the product of the number of syllables recalled and the average rate of information, is constant for all passages." [p. 261, 1] The data verified the first hypothesis, but not the second. For the first four passages the total amount of information learned was constant, but it dropped in passage 5 and even more so in passage 6. The breaking point was between 1.5 and 2 bits/syllable. This result simply means that the subjects were unable to memorize enough syllables to keep the information score high when the information per syllable was very low. Both these findings are in conformity with those of Miller and Selfridge above.

8.2 Operant Conditioning

Frick and Miller [19] have reported an application of their earlier ideas for the measurement of stereotypic behavior [65] to the operant conditioning of rats in a Skinner box. Two responses were observed: approach to food (A) and bar pressing (B). "Instead of the usual analysis in terms of

the rate of responding to the bar, the results are analyzed here in terms of the patterns of responses." [p. 21, 19] Three experimental phases were considered separately in the analysis: a) behavior prior to conditioning (operant level), b) conditioning behavior, and c) extinction behavior. During phase b a total of 300 reinforcements was applied.

In all phases the behavior was recorded as sequences of A's and B's, and the uncertainties - in terms of the index of behavioral stereotypy- were computed. It was found that 'intersymbol' influences did not extend appreciably beyond two symbols, and the value of the uncertainty in phase a was 0.408 for two symbols. Such a high value when there has been no conditioning is a consequence of the fact that such a sequence as AAAA had a probability of 0.732 of occurring; indeed, the behavior of the rats was more stereotyped before conditioning than after. "The training-period did not introduce order into randomness, but rather caused the animal to abandon one well organized pattern of behavior for another. This needs some qualification. The lower stereotypy after conditioning appears when we consider only the temporal order; when we try to predict which response comes next. If we tried to predict also when the next response would occur and how long it would last, then the conditioned behavior would look less random than the pre-conditioned behavior." [p. 25, 19]

Another simple way the data may be described is as points in a two-dimensional plot of $p(B|B)$ vs $p(A|A)$. In phase a of the experiment the rats were approximately at the point (0.9, 0.75). This high perseveration is, in large part, simply a reflection of the topography of the Skinner box, as

can be seen from the fact that 96 per cent of the responses separated by less than 10 seconds are of the form AA and BB, while this is reduced to 52 per cent for responses separated by more than 80 seconds.

During conditioning, phase b, the rats initially move down the plot and then curve slowly over to an equilibrium point of about (0.1, 0.4), as shown in figure 6. During the extinction period the movement of a rat in this space is not very clear.

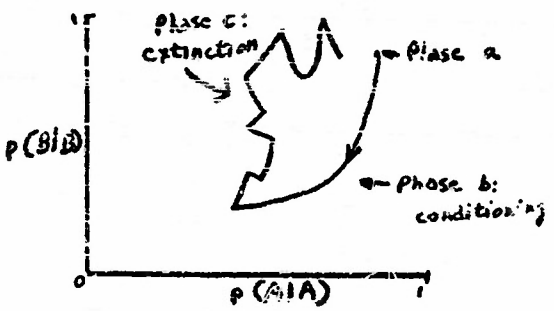

Fig. 6

There appears to be an initial tendency toward the center (0.5, 0.5) of the plot, or random behavior, but there is considerable random variation over a large portion of the plot. Over a 36 hour period there is a drift toward the initial resting point, but no stability is achieved in that period comparable to that prior to conditioning. It was not determinable from this data how long it takes for the effects of reinforcement to wear off. As in phase a, there is little difference between the uncertainty determined from two successive responses and from more than two, and after some extinction there is little or no difference in the index based on a single response and that based on successive pairs of responses.

"The data presented and analyzed [in this paper] do not provide any startling new insights into operant conditioning. Most of the conclusions seem perfectly reasonable and obvious to anyone who has worked with rats in a similar situation and observed their general behavior closely. The impressive feature of such an analysis is the extent to which the qualitative

aspects of the behavior can be incorporated into a completely quantitative account." [p. 35, 19]

8.3 Intelligibility

The data on the effects of sequential dependencies on intelligibility are less detailed than for learning, but there is an experiment by Miller, Heise, and Lichten [69] in which certain gross effects were examined. They explored the effects of three different contexts on intelligibility, namely: the test item is known to be one of a small vocabulary of possible items, the test item is imbedded in either a word or a sentence, and the test item is known to be a repetition of the preceding item. The materials used were digits, words in sentences, and nonsense syllables, and it was found that intelligibility decreased in that order. Further, the intelligibility of monosyllables, isolated words, and words in sentences was found to increase in each case as the domain of possible items was decreased. Only a very slight increase in intelligibility resulted from the knowledge that the item was a repetition of the preceding one. "The results indicate that far more improvement in communication is possible by standardizing procedures and vocabulary than by merely repeating all messages one or two times." [p. 335, 69] This conclusion seems to confirm the military practice of using standardized languages when conditions are adverse, as in air traffic control (see section II.2.4).

8.4 Perception

Hake and Hyman [31] raised the question of just how well and in what way people perceive sequential dependencies which are built into a set of stimuli, and they chose to summarize their results in terms of certain conditional uncertainties - entropies - of the subject responses. The experiment was divided into four series of runs. Each run consists of 240 presentations of one or the other of two symbols (H and V), and these presentations were generated according to the following probabilities and conditional probabilities:

	Series 1	2	3	4
p(H)	.50	.50	.75	.75
p(H|H)	.50	.80	.75	.90
p(V)	.50	.50	.25	.25
p(V|V)	.50	.80	.25	.70

Prior to each presentation, the subjects were required to predict, or guess, which symbol would occur. The problem of analysis is to determine how accurately we can predict his guess provided we know certain past events such as his guesses and the symbols which actually occurred. For the last 120 trials the following conditional entropies were examined: the entropy of the guess y when only the distribution of y is known -H(y), the entropy of y when the distribution of y and the previous guess are known $-H_y(y)$,

the entropy of y when the distribution of y, the previous guess, and the previous occurrence are known - $H_{xy}(y)$, the entropy of y when the distribution of y and the previous occurrence are known - $H_x(y)$, and the analogues of each of the last three for the two preceding trials, instead of just one. These data are summarized:

	Series 1	2	3	4
H(y)	1.00	1.00	.76	.80
$H_y(y)$	1.00	.83	.72	.75
$H_x(y)$	.99	.69	.74	.70
$H_{xy}(y)$	.98	.54	.68	.55
$H_{y,y}(y)$	1.00	.83	.72	.73
$H_{x,x}(y)$	.98	.55	.70	.56
$H_{xy,xy}(y)$	.95	.52	.66	.55

It is clear that the best prediction of the subject's guess, i.e., the lowest entropy, is obtained when both his guesses and the actual occurrences on the two preceding trials are known, but a knowledge of his guess and the actual occurrence on the single preceding trial yields a prediction which is nearly as good, and knowledge of only the occurrence on the two preceding trials is only slightly worse. It thus follows that a subject responds not only to the actual events which occurred but also to his predictions about them. This can be made quite apparent by computing the probability of a guess of H when on the preceding trial a <u>correct</u> guess of H was made. For

series one this conditional probability is about 0.5, but for the other three series it rises over trials and from trial 100 on it remains approximately constant with a value of 0.9. When the probability of an H guess following two successive correct H guesses is plotted, the curves rise more rapidly, and even in series one there is a rise from 0.5 to about 0.75.

"We conclude from our evidence that Ss do not, in fact, perceive the probability rules by which the series of events was generated. They do perceive, instead, those short sequences of events which precede each prediction, which can be discriminated from other possible sequences, and which are found to provide some information about the future behavior of the symbol series. There are several interesting conclusions which we can make about the way in which Ss perceive these previous events.

"1. All combinations of possible previous events were not discriminated with equal ease. Some previous events, especially homogeneous runs of the same symbol, were more easily discriminated and consistently responded to than were others.

"2. The previous events to which our Ss responded on each trial included more than just the symbols which had been appearing. They included also the previous predictions of Ss and the degree of correspondence between their predictions and the symbols which appeared on previous trials.

"3. There was considerable agreement among our Ss as to <u>when</u> a particular symbol should be predicted. They tended to respond to some similar or identical previous events in the same way, no matter which series they were predicting..." [p. 72, 31]

9. Immediate Recall of Sets of Independent Selections

The subject of this section is closely related to that of II.8.1; it was not included there since the main emphasis of that section was on the effects of inter-symbol dependencies on immediate recall, whereas here we shall examine the effects of message length and the bits/symbol when there are no dependencies among symbols. Pollack [77] prepared messages of from 4 to 24 symbols from sets of 2, 4, 8, 16, and 30 equi-probable English consonants and numerals. These were read in a uniform manner to subjects who were told in advance both the set of symbols and the message length, and they were required to reproduce them as accurately as possible. When an error was made, the subject was requested to guess as many times as was necessary to obtain the correct response. In one version of the experiment, reading rates were varied, but "Rate of presentation of stimulus materials (over the range considered) appears as a variable with little significance for immediate recall under the conditions considered here." [p. 13, 77, II]

The data show that the error entropy per message unit increases both with message length and with an increase in bits/symbol, but that for a message of given length the percentage of presented information which is lost is approximately independent of the number of bits/symbol. This percentage is, however, an increasing function of the length of the message. The error entropy increased in such a manner that the total information transmitted increased as the message length was increased from 4 symbols to about 10, it remained roughly constant in the range of 10 to 16 or 18

symbols per message, and it decreased for longer messages. The curves are displaced upward with an increase in bits/symbol, but they are of remarkably similar shape. "The main generalization is that one cannot obtain simultaneously both minimum information loss and maximum information gain by simply varying either the length of a message or the number of possible alternatives per message-unit." "These relations stem from the fact that the percentage of the information presented that is lost or gained is independent of the number of alternatives per unit and is simply a function of the length of the message." [p. 12, 77, I]

It is useful to transform these data into plots of error entropy and information transmitted vs total informational input. It is then found that for a fixed input, the error entropy is smaller and the information transmitted is larger the larger the number of bits/symbol. Thus, as Pollack points out, if one is interested in the optimal encoding characteristics for messages of fixed length, there are two answers, depending on whether a high error count is tolerable or not. If, however, the question is what are the optimal encoding characteristics (for immediate recall) for messages of fixed informational content, then the answer is unequivocal: short messages with a large number of alternatives for each message unit.

In parts III and IV of his report, Pollack systematically studied the error behavior of his subjects. First, his data confirm the familiar finding of this type of experiment that the subjects are most uncertain about the middle portion of the message. For messages of length 7, the relative uncertainty of the middle symbols is slightly higher than the end uncertainty,

but it never exceeds .30. However, for messages of length 24, there is a broad plateau in the middle of the message which has a relative uncertainty of about .80. The broadness of this plateau Pollack attributes to the great sensitivity of the information measure to errors. He notes that this uncertainty curve alters its character with increasing message length: for short messages it is positively skewed and for long ones it is negatively skewed.

In the fourth part of the report, he established the conclusion that there is still information transmitted (as compared with chance responses) by the subjects on the second and third guesses following an incorrect response. "In general, the additional information recovered per message increases as the degree of analysis of the multiple response data becomes more exhaustive. Stated otherwise, we recover more information from the distribution of responses if we utilize the first response following the initial incorrect reproduction, still more if we utilize the first and second responses following the initial incorrect reproduction, and still more if we utilize the first through the third responses following the initial incorrect reproduction. The magnitude of the information recovered increases as the number of alternatives per message-unit increases and is, roughly, independent of message-length (for messages greater than 7 units in length)." [p. 8, 77, IV] As would be expected, this effect is a decreasing one, but the decrease is less rapid with larger numbers of alternatives per message-unit.

10. Concept Formation

Let there be eight objects which are triangles or circles, large or small, and black or red. We may attempt to convey a concept, such as red triangle, to a subject by showing him the objects one at a time and stating whether or not they are examples of the desired concept. A positive instance of the concept red triangle is large red triangle, whereas small black triangle or large red circle are negative instances. Such experiments in concept learning have long been performed, and the conclusion has been drawn that negative instances are of little value in learning the correct concept. Hovland [37], however, has raised a question about this conclusion - a question which stems from an information analysis of the situation. "What is not clear ... is whether the ineffectiveness of negative instances is primarily attributable to their low value as carriers of information, or whether it is primarily due to the difficulty of assimilating the information which they do convey." [p. 461, 37]

Certainly it is clear from the above example that positive and negative instances do not transmit the same information, since only two positive ones are required to specify the concept, as compared with six negative. It is, of course, possible to design a situation where the negative instances carry as much or more information as the positive ones. For certain simple general situations, of which the above example is illustrative, Hovland has given formulas for the total number of positive and negative instances required to specify the concept. In an experimental paper, he and Weiss [38] examined the effect of positive and negative

instances when both the number of instances and the amount of information are held constant, and they conclude that even so the negative instances do not contribute as effectively to learning. "At the same time the data disprove the generalization often cited that negative instances have no value in the learning of concepts. Under appropriate conditions over half of the Ss were able to reach the correct solution solely on the basis of negative instances." [p. 181, 28]

Bendig [4] conducted an experiment which is closely related to concept formation, namely, the identification of a concept after the manner of the game '20 questions.' In the experiment, four questions were employed to isolate an animal topic. One experimenter asked the questions in fixed order of another who answered 'yes' or 'no' according to the topic. Following each question, the subjects were required to guess the concept. The information transmitted by each question was calculated, and theoretically each should have conveyed one bit, but in actuality 0.03, 0.91, 0.21, and 0.78 bits were transmitted. The central conclusion seemed to be that the third question was unfortunately phrased, since answers to it failed to convey much information.

II. Paired Associates Learning

In this section, we shall consider a learning situation where one class of objects - usually words - known as 'responses' have been placed in one-to-one correspondence with another class of objects known as 'stimuli.' Initially, the subject knows nothing of the pairing and he can only guess at the appropriate response to a given stimulus: if he is correct,

he is told this, if not, he is told the correct response. After a number of repetitions, R, of the stimulus class, the subject begins to learn the correct pairing, and he obtains a certain number of correct bonds, say C, out of the total of N. The function C(R) is known as his 'learning curve' for the paired associates. Several theories, and formulae, for this learning phenomenon have been put forth which are summarized by Rogers [8b] in a thesis in which he introduces a new learning theory based in part on information theory.

He makes two central assumptions. First, he supposes that the uncertainty which a subject has with respect to the stimulus class after R repetitions of the stimulus class is a function of R alone. In particular, he supposes that it is constant - U_{ok} - for the first b repetitions, where b is a 'set' parameter which tells when the learning begins, and that from b on it is a linear function of R, i.e.,

$$U_R = U_{ok} - a(R-b) \quad \text{for } R \geq b.$$

Second, let B be the total number of bonds which the subject knows after R repetitions, which Rogers shows is one less than the expected value of the observable C. Let k be a stimulus not among the B that are known and let i be any response which is not associated with one of the B known stimuli, then he supposes that the probability that i is the response when k is given is $1/(N-B)$. In other words, the subject is assumed to distribute his response choices without preference over all the available response elements.

From this second assumption, it is not difficult to obtain an

expression for the uncertainty in terms of N and B. Equating this to the assumed expression in terms of R gives an equation between B and R, and so between C and R. This may be solved for C:

$$C = \begin{cases} (N-1)\left[1 - e^{-\delta a(R-b)}\right] + 1, & \text{for } R \geq b \\ 1, & \text{for } R < b \end{cases}$$

where $\delta = \log_2 e$. It has long been noted that many learning data are approximately fit by such an exponential learning curve, though in general this has been an empirical observation which was not deduced from other assumptions.

To test the merits of this theory, Rogers drew certain conclusions from it which could be confronted by data, and these conclusions were sustained by his data. Three experiments of a similar type were performed. 1) Correlated Structure. Stimuli - playing cards having two easily recognized dimensions, suits and denominations, were associated with nonsense syllables of the form consonant-vowel-consonant in a correlated manner. The first letter always corresponded to the denomination and the last to the suit. 2) Unstructured. Pictures of diverse household objects were paired in an arbitrary manner with nonsense syllables. 3) Uncorrelated Structured. The same materials as in 1 were used (so both the stimulus class and the response class were structured) but there was no systematic relation in the pairing between the stimulus class and the response class. He then examined what two classical theories of learning - Gestalt and the

transfer theory of meaning -- and his own information theory of learning predicted as to the learning rates in these three cases. Gestalt theory, according to his interpretation, ranks them 1, 3, 2 in order of increasing difficulty, transfer theory gives an ordering of 1, 2, 3, while information theory predicts that 1 and 2 should be equally easy and 3 more difficult. His data are consistent with only the last prediction.

Attempts to fit the learning curve to the data were for the most part successful, although one can note a consistent 'S' character to the data, which, of course, the exponential does not possess. He points out that if the linear assumption were replaced by an appropriate non-linear one, once could easily produce a learning curve with an 'S' shape - or, we might add, with practically any other shape, for that matter.

Appendix: The Continuous Theory

Much communication can best be thought of as the transmission of a continuous signal and not as a sequence of temporally ordered selections from a finite set of possible elements. For the most part, as we have seen, the continuous theory has been of little importance in behavioral applications, though it is of considerable importance in electrical ones. We shall, therefore, only sketch the theory briefly. Our presentation follows Shannon's [87] closely.

A.1 The Continuous Source

A source is said to be continuous if, in effect, it makes but one selection from a continuum of elements; specifically, if it chooses one number from the set of all real numbers. We shall suppose that this selection is characterized by the probability distribution $p(x)$ over the real numbers x. Since p is a distribution, $\int_{-\infty}^{\infty} p(x)dx = 1$, and furthermore for any $\varepsilon > 0$, no matter how small, we can find finite a and b such that $1 - \varepsilon < \int_a^b p(x)dx \leq 1$. Now, for such a and b we may divide the interval from a to b into n equal intervals, and we can treat each of the intervals as an element from a finite set, with probability $\int_{x_i}^{x_{i+1}} p(x)dx$ of being selected.

All the continuum not in a to b is an $n+1$st element with probability

$1 = \int_a^b p(x)dx$. Thus we have approximated the continuous source by a discrete one and for each n we can compute a corresponding entropy H_n. As we let n approach infinity, the approximation is better and better, but unfortunately H_n also approaches infinity. This, of course, is reasonable considering the basis of the discrete entropy concept, but that does not make the approach any more satisfactory as a way to compare continuous sources.

In such situations it is very often the case that the difference between the quantity desired and another quantity which tends to infinity with n will itself tend to a finite limit. If this second quantity can be chosen to be the same for all sources, then the resulting differences are perfectly acceptable comparators for the continuous source. As before, we choose a and b and we divide the interval from a to b into n equal intervals. Each of these intervals is of length $\Delta x = (b-a)/n$. Whereas before we tried to generalize

$$-\sum_{i=1}^{n} p(x_i)\Delta x \log_2 p(x_i)\,\Delta x$$

and got into trouble, we now examine

$$\log_2 \Delta x = \sum_{i=1}^{n} p(x_i)\,\Delta \log_2 p(x_i)\,\Delta x.$$

It is not difficult to show that

$$\lim_{\substack{b\to\infty \\ a\to-\infty}} \lim_{\substack{n\to\infty \\ \Delta x\to 0}} \left[\log_2 \Delta x - \sum_{i=1}^{n} p(x_i)\,\Delta x \log_2 p(x_i)\,\Delta x\right]$$

$$= - \int_{-\infty}^{\infty} p(x) \log_2 p(x).$$

This quantity, which is denoted H(x), is called the entropy of a continuous source. It is well to keep in mind that the continuous entropy is not an exact analogue of the discrete entropy, and so certain differences in properties may be expected. The surprising thing is how many of the results are independent of the base-line from which the discrete entropy is measured.

If there are two arguments x and y to the distribution (as in the case of noise), the joint and conditional entropies are defined by

$$H(x,y) = - \int\int p(x,y) \log_2 p(x,y)\, dxdy,$$

$$H_x(y) = - \int\int p(x,y) \log_2 \frac{p(x,y)}{p(x)}\, dxdy$$

$$H_y(x) = - \int\int p(x,y) \log_2 \frac{p(x,y)}{p(y)}\, dxdy,$$

where

$$p(x) = \int p(x,y)dy$$

$$p(y) = \int p(x,y)dx.$$

Many of the theorems of the discrete case carry over - usually quite directly - to the continuous case, but in addition there are certain new theorems which rest heavily on the existence of a coordinate system. We list some of the more important ones, of which the first is familiar and the other four are new.

1. $H(x,y) \leq H(x) + H(y)$,

$$H(x,y) = H(x) + H_x(y) = H(y) + H_y(x),$$

$$H_x(y) \leq H(y).$$

2. If $p(x) = 0$ except on an interval of length v, then $H(x)$ is a maximum ($= \log_2 v$) when $p(x) = 1/v$ for x in the interval.

3. Of the class of all continuous one-dimensional distributions with variance σ^2, the normal, or Gaussian, is the one having maximum entropy. The value of the maximum is $\log_2 2(\pi)^{1/2}\sigma$.

4. Of the class of all continuous one-dimensional distributions with mean $a > 0$ and with $p(x) = 0$ for $x \leq 0$, the exponential is the one having maximum entropy. The value of the maximum is $\log_2 ea$.

5. Unlike the discrete case, in which entropy measures the randomness in an absolute way, the continuous entropy is a measure which is relative to a coordinate system. If the coordinate system is changed, the entropy is changed. This is not serious, however, since both the channel capacity and the rate of information transfer depend on the difference of two entropies, and so they are invariant under coordinate transformation. Reich [83] states that he has shown that the definition of information rate used by Shannon is the only one of a broad class of possible definitions which is invariant under coordinate transformation.

A.2 The Channel Capacity

As in the discrete noisy case, the channel capacity C is defined to be the maximum rate of transmission $R = H(x) - H_y(x)$ obtained by considering

all possible distributions. This is easily shown to be

$$C = \lim_{T\to\infty} \max_{p(x)} \frac{1}{T} \int\int p(x,y) \log_2 \frac{p(x,y)}{p(x)\,p(y)}\, dx\, dy.$$

One particularly important case in applications is that in which the noise is simply added to the signal and is independent of it. In that case the entropy of the noise can be computed. If we denote it by H(n), then

$$C = \max_{p(x)} H(y) - H(n).$$

Of course, if there are restraints on the class of admissible signals, the maximization is taken subject to these restraints.

A simple, but very important, electrical application of the above theorem is to the case of a channel which has a bandwidth of W cycles per second (e.g., a telephone which will pass from 500 to 3,500 cycles per second has a bandwidth of 3,000 cycles per second), in which the transmitter has an average power output of P and the noise is white thermal noise(i.e., all frequencies are equally represented) of average power N. In this case the channel capacity in bits per second is

$$C = W \log_2 \left(1 + \frac{P}{N}\right).$$

4.3 Rate of Transmission

"In the case of a discrete source of information we were able to determine a definite rate of generating information, namely the entropy of the underlying stochastic process. With a continuous source the situation

is considerably more involved. In the first place a continuously variable quantity can assume an infinite number of values and requires, therefore, an infinite number of binary digits for exact specification. This means that to transmit the output of a continuous source with exact recovery at the receiving point requires, in general, a channel of infinite capacity (in bits per second). Since, ordinarily, channels have a certain amount of noise, and therefore a finite capacity, exact transmission is impossible.

"This, however, evades the real issue. Practically, we are not interested in exact transmission when we have a continuous source, but only in transmission to within a certain tolerance. The question is, can we assign a definite rate to a continuous source when we require only a certain fidelity of recovery, measured in a suitable way. Of course, as the fidelity requirements are increased the rate will increase. It will be shown that we can, in very general cases, define a rate, having the property that it is possible, by properly encoding the information, to transmit it over a channel whose capacity is equal to the rate in question, and satisfy the fidelity requirements. A channel of smaller capacity is insufficient." [p. 74, 80]

The noise characteristic of the whole system is, as before, given by a distribution $p(x,y)$ which states the distribution that the signal y is received when x is sent. The fidelity of the system is, roughly, an evaluation of how different y is on the average from x. It is assumed to be a function of the noise, that is, if it is measured by a real number it can be written in the form $v(p(x,y))$. Under quite broad conditions, which

we shall not attempt to state here (see [87]), it can be shown that v can be represented as:

$$v(p(x,y)) = \int\int p(x,y)\, \rho(x,y)\, dx\, dy.$$

The real-valued function $\rho(x,y)$ is essentially a measure of the difference between x and y and in computing the fidelity it is weighted according to the probability density of the joint occurrence of x and y. It may be illuminating to consider two very common electrical criteria of fidelity. The first is the root-mean-square criterion, namely,

$$\rho(x,y) = \frac{1}{T} \int_0^T [x(t) - y(t)]^2\, dt,$$

and the second is the absolute error criterion, namely,

$$\rho(x,y) = \frac{1}{T} \int_0^T |x(t) - y(t)|\, dt.$$

Now, the rate R of generating information corresponding to a given quality of reproduction (fidelity) v is defined to be the minimum R which is obtained by varying $p(y|x)$ with v held constant, i.e.,

$$R = \min_{p(y|x)} \int\int p(x,y)\, \log_2 \frac{p(x,y)}{p(x)\, p(y)}\, dx\, dy$$

subject to

$$v = \int\int p(x,y)\, \rho(x,y)\, dx\, dy.$$

With this definition, and with that of channel capacity given in section A.2, it can be shown that if a source has a rate R for a valuation of fidelity v, then it is possible to encode the output of the source and to

transmit it over a channel with capacity C in such a manner that the fidelity is arbitrarily near v if, and only if, $R \leq C$. This is the fundamental theorem for the transmission of information in the continuous case.

Bibliography

The following group of papers and books which were examined in the preparation of this report include the central works on the theory of information and all of the works which we have been able to find (as of early 1954) concerned with its application to psychology. The bibliography prepared by Stumpers [95, 96] is more general than ours in that it covers the whole area of Cybernetics and the applications of information theory in engineering and in the several behavioral sciences (as of early 1953), but it is not so complete as ours for psychological applications.

1. Aborn, M. and Rubenstein, H., "Information Theory and Immediate Recall," Journal of Experimental Psychology, 44, 1952, 260-266.

2. Bell, D. A., "The 'Internal Information' of English Words," Communication Theory (ed. Willis Jackson) Academic Press, Inc., New York, 1953, 385-391.

3. Bendig, A. W. and Hughes, J. B., "Effect of Amount of Verbal Anchoring and Number of Rating-Scale Categories Upon Transmitted Information," Journal of Experimental Psychology, 46, 1953, 87-90.

4. Bendig, A. W., "Twenty Questions: An Information Analysis," Journal of Experimental Psychology, 46, 1953, 345-348.

5. Blachman, N. M., "Minimum-Cost Encoding of Information," Transactions of the Institute of Radio Engineers, Professional Group on Information Theory, 3, 1954, 139-149.

6. Carnap, R. and Bar-Hillel, Y., An Outline of a Theory of Semantic Information, Research Laboratory of Electronics Technical Report, 247, M.I.T., 1952.

7. Cherry, E. C., "A History of the Theory of Information," Proceedings of the Institution of Electrical Engineers, III, 98, 1951, 383-393.

8. _____, "A History of the Theory of Information," Transactions of the Institute of Radio Engineers, Professional Group on Information Theory, 1, 1953, 22-43.

9. Christie, L. S. and Luce, R. D., Suggestions for the Analysis of Reaction Times and Simple Choice Behavior, ditteed paper, 1953.

10. Crossman, E. R. F. W., "Entropy and Choice Time: the Effect of Frequency Unbalance on Choice Response," Quarterly Journal of Experimental Psychology, 5, 1953, 41-52.

11. Dewey, G., Relative Frequency of English Speech Sounds, Harvard University Press, Cambridge, 1923.

12. Delansky, Ludmilla and Delansky, M. P., Table of $\log 2 \frac{1}{p}$, $P \log 2 \frac{1}{p}$, and $P \log 2 \frac{1}{p} + (1-P) \log 2 \frac{1}{1-p}$, Technical Report 227, Research Laboratory of Electronics, M.I.T., 1952.

13. Elias, Peter, "A Note on Autocorrelation and Entropy," Proceeding of the Institute of Radio Engineers, 39, 1951, 839.

14. Fano, R. M., The Transmission of Information, Research Laboratory of Electronics Technical Report 65, M.I.T., 1949.

15. _____, The Transmission of Information - II, Research Laboratory of Electronics Technical Report 149, M.I.T., 1950.

16. _____, "The Information Theory Point of View in Speech Communication," Journal of the Acoustical Society of America, 22, 1950, 691-696.

17. _____, Information Theory, Past, Present and Future, ditto'd paper, M.I.T., 1951.

18. Felton, W. W., Fritz, E., and Grier, G. W., Jr., Communication Measurements at the Langley Air Force Base, Human Resources Research Laboratory Report No. 31, 1951.

19. Frick, F. C. and Miller, G. A., "A Statistical Description of Operant Conditioning," American Journal of Psychology, 64, 1951, 20-36.

20. Frick, F. C. and Sumby, W. H., "Control Tower Language," Journal of the Acoustical Society of America, 24, 1952, 595-597.

21. Gabor, D., "Theory of Communication," Journal of the Institution of Electrical Engineers, 93, III, 1946, 429-456.

22. _____, "New Possibilities in Speech Transmission," Journal of the Institution of Electrical Engineers, 94, III, 1947, 369-390.

23. _____, Lectures on Communication Theory, Research Laboratory of Electronics Technical Report 238, M.I.T., 1952.

24. _____, "Communication Theory, Past, Present, and Prospective," Transactions of the Institute of Radio Engineers, Professional Group on Information Theory, 1, 1953, 2-4.

25. _____, "A Summary of Communication Theory," Communication Theory (ed. Willis Jackson), Academic Press, Inc., New York, 1953, 1-23.

26. _____, "Communication Theory and Physics," Transactions of the Institute of Radio Engineers, Professional Group on Information Theory, 1, 1953, 48-59.

27. Garner, W. R. and Hake, H. W., "The Amount of Information in Absolute Judgments," Psychological Review, 58, 1951, 446-459.

28. Garner, W. R., "An Informational Analysis of Absolute Judgments of Loudness," Journal of Experimental Psychology, 46, 1953, 373-380.

29. Goldman, S., Information Theory, Prentice-Hall, New York, 1953.

30. Hake, H. W. and Garner, W. R., "The Effect of Presenting Various Numbers of Discrete Steps on Scale Reading Accuracy," Journal of Experimental Psychology, 42, 1951, 358-366.

31. Hake, H. W. and Hyman, R., "Perception of the Statistical Structure of a Random Series of Binary Symbols," Journal of Experimental Psychology, 45, 1953, 64-74.

32. Halsey, R. M. and Chapanis, A., "On the Number of Absolutely Identifiable Spectral Hues," Journal of the Optical Society of America, 42, 1952, 1057-1058.

33. Hartley, R. V. L., "Transmission of Information," Bell System Technical Journal, 7, 1928, 535-563.

34. Hick, W. E., "On the Rate of Gain of Information," Quarterly Journal of Experimental Psychology, 4, 1952, 11-26.

35. _____, "Why the Human Operator?" Transactions of the Society of Instrument Technology, 4, 1952, 67-77.

36. _____, "Information Theory in Psychology," Transactions of the Institute of Radio Engineers, Professional Group on Information Theory, 1, 1953, 130-133.

37. Hovland, C. I., "A 'Communication Analysis' of Concept Learning," Psychological Review, 59, 1952, 461-472.

38. Hovland, C. I. and Weiss, W., "Transmission of Information Concerning Concepts Through Positive and Negative Instances," Journal of Experimental Psychology, 45, 1953, 175-182.

39. Howes, D. H., The Definition and Measurement of Word Probability, Ph.D. Thesis, Harvard University, 1950.

40. Howes, D. H. and Solomon, R. L., "Visual Duration Threshold as a Function of Word Probability," Journal of Experimental Psychology, 41, 1951, 401-410.

41. Hyman, R., "Stimulus Information as a Determinant of Reaction Time," Journal of Experimental Psychology, 45, 1953, 188-196.

42. Jackson, Willis (Editor), "Report of Proceedings, Symposium on Information Theory, London, 1950," Transaction of the Institute of Radio Engineers Professional Group on Information Theory, 1, 1953.

43. _____, Communication Theory, Academic Press Inc., New York, 1953.

44. Jacobson, H., "The Informational Capacity of the Human Ear," Science, 112, 1950, 143-144.

45. _____, "Information and the Human Ear," Journal of the Acoustical Society of America, 23, 1951, 463-471.

46. _____, "The Informational Capacity of the Human Eye," Science, 113, 1951, 292-293.

47. King-Wilson, Patricia and Jenkins, J. J., Visual Duration Threshold as a Function of Word Frequency: A Replication, The Role of Language in Behavior, Technical Report Number 6, University of Minnesota, Contract No. N8 onr-66216.

48. Klemmer, E. T. and Frick, F. C. "Assimilation of Information from Dot and Matrix Patterns," Journal of Experimental Psychology, 45, 1953, 15-19.

49. Klemmer, E. T. and Muller, P. F., Jr., The Rate of Handling Information: Key Pressing Responses to Light Patterns, HFORL memo Report No. 34, 1953.

50. Krulee, G. K. and Sinclair, E. J., "Some Behavioral Implications of Information Theory," Report 4119, Naval Research Laboratory, Washington, D.C., 1953, 11 pp.

51. Kullback, S. and Leibler, R. A., "On Information and Sufficiency," Annals of Mathematical Statistics, 22, 1951, 79-86.

52. Kullback, S., "An Application of Information Theory to Multivariate Analysis," Annals of Mathematical Statistics, 23, 1952, 88-102.

53. Licklider, J.C. R. and Miller, G. A., "The Perception of Speech," Handbook of Experimental Psychology (S. S. Stevens, editor), John Wiley and Sons, 1951, 1040-1074.

54. MacKay, D. M., "The Nomenclature of Information Theory," Cybernetics (ed. Heinz von Foerster) Josiah Macy, Jr. Foundation, New York, 1951, 222-233; and Transactions of the Institute of Radio Engineers, Professional Group on Information Theory 1, 1953, 9-21.

55. _____, "Quantal Aspects of Scientific Information," Philosophical Magazine (series 7), 41, 1950, 289-311; and Transactions of the Institute of Radio Engineers, Professional Group on Information Theory, 1, 1953, 60-80.

56. _____, "In Search of Basic Symbols," Cybernetics (ed. Heinz von Foerster), Josiah Macy, Jr. Foundation, New York, 1951, 181-221.

57. Mandelbrot, Benoît, "Contribution à la Théorie Mathématique Des Jeux de Communication," Publications de l'Institut de Statistique de l'Université de Paris, 2, 1953, 1-124.

58. _____, "An Informational Theory of the Statistical Structure of Language," Communication Theory (ed. Willis Jackson), 1953, Academic Press, New York, 486-502.

59. _____, "Structure Formelle des Textes et Communication Deux Études," Word, 10, 1954, 1-27.

60. _____, "Simple Games of Strategy Occurring in Communication Through Natural Languages," Transactions of the Institute of Radio Engineers, Professional Group on Information Theory, 3, 1954, 124-137.

61. McGill, W. J., Multivariate Transmission of Information and its Relation to Analysis of Variance, Report No. 32, Human Factors Operations Research Laboratory, M.I.T., 1953.

62. _____, "Multivariate Information Transmission," Psychometrika, to be published; and Research Laboratory of Electronics and Lincoln Laboratory Technical Memorandum No. 48, M.I.T., 1953, 17 pp.

63. McMillan, Brockway, "The Basic Theorems of Information Theory," The Annals of Mathematical Statistics, 24, 1953, 196-219.

64. Merkel, J., "Die Zeitlichen Verhaltnisse der Willensthatigkeit," Philos. St., 2, 1885, 73-127.

65. Miller, G. A. and Frick, F. C., "Statistical Behavioristics and Sequences of Responses," Psychological Reviews, 56, 1949, 311-324.

66. Miller, G. A. and Selfridge, J. A., "Verbal Context and the Recall of Meaningful Material," American Journal of Psychology, 63, 1950, 176-185.

67. Miller, G. A., "Speech and Language," Handbook of Experimental Psychology (S. S. Stevens, editor), John Wiley and Sons, 1951, 789-810.

68. _____, Language and Communication, McGraw-Hill, New York, 1951.

69. Miller, G. A. and Heise, G. A., and Lichten, W., "The Intelligibility of Speech as a Function of the Context of the Test Materials," Journal of Experimental Psychology, 41, 1951, 329-335.

70. Miller, G. A., A Note on the Sampling Distribution of the Shannon-Wiener Measure of Information, Unpublished paper, 1952.

71. _____, "What is Information Measurement?" American Psychologist, 8, 1953, 3-11.

72. _____, "Communication," Annual Review of Psychology, 5, (Stone, C. P. and McNemar, Q., editors). Annual Reviews, Inc., Stanford, 1954, 401-420.

73. Miller, G. A. and Madow, W. G., On the Maximum Likelihood Estimates of the Shannon-Wiener Measure of Information, (in preparation), 1954.

74. Newman, E. B., "Computational Methods Useful in Analyzing Series of Binary Data," American Journal of Psychology, 64, 1951, 252-262.

75. Newman, E. B. and Gerstman, C. J., "A New Method for Analyzing Printed English," J. Exp. Psychology, 44, 1952, 114-125.

76. Pollack, Irwin, "Information of Elementary Auditory Displays," Journal of the Acoustical Society of America, 24, 1952, 745-750.

77. _____, The Assimilation of Sequentially-Encoded Information. 1. Methodology and an Illustrative Experiment. 2. Effect of Rate of Information Presentation. 3. Serial Position Analysis. 4. The Informational Contribution of "Wrong" Responses. Human Resources Research Laboratories Technical Report No. 25, Washington, 1952.

78. Pratt, Fletcher, Secret and Urgent, Blue Ribbon Books, Garden City, 1942.

79. Proceedings of the London Symposium on Information Theory, 1950, See Jackson [42]

80. Proceedings of the London Symposium on Information Theory, 1952, See Jackson [43]

81. Quastler, Henry (editor), Essays on the Use of Information Theory in Biology, University of Illinois Press, Urbana, Ill., 1953.

82. Quastler, Henry and Wolff, W. J., Human Performance in Information Transmission Part One: Simple Sequential Routine Tasks, Unpublished mss., 1953.

83. Reich, E. "Definition of Information," Proceedings of the Institute of Radio Engineers, 39, 1951, 790.

84. Rogers, M. S., An Application of Information Theory to the Problem of the Relationship Between Meaningfulness of Material and Performance in a Learning Situation, Ph.D. Thesis, Princeton University, 1952, mimeographed.

85. Rogers, M. S. and Green, B. F., The Moments of Sample Information when the Alternatives are Equally Likely, Mimeographed paper, Psychological Laboratories, Harvard University, 1954.

86. _____, Tables of the Mean and Variance of Sample Information when the Alternatives are Equally Likely, Mimeographed, Psychological Laboratories, Harvard University, Cambridge.

87. Shannon, C. E., "A Mathematical Theory of Communication," Bell System Technical Journal, 27, 1948, 379-423 and 623-656.

88. Shannon, C. E. and Weaver, Warren, The Mathematical Theory of Communication, University of Illinois Press, Urbana, 1949.

89. Shannon, C. E., "Communication Theory of Secrecy Systems," Bell System Technical Journal, 28, 1949, 656-715.

90. _____, "The Redundancy of English," Cybernetics (ed. Heinz von Foerster) Josiah Macy, Jr. Foundation, New York, 1950, 123-158.

91. _____, "Prediction and Entropy of Printed English," Bell System Technical Journal, 30, 1951, 50-64.

92. _____, "Communication Theory, Exposition of Fundamentals," Transactions of the Institute of Radio Engineers, Professional Group on Information Theory, 1, 1953, 44-47.

93. _____, "General Treatment of the Problem of Coding," Transactions of the Institute of Radio Engineers, Professional Group on Information Theory, 1, 1953, 102-104.

94. _____, "The Lattice Theory [illegible] formation," Trans[illegible] of Radio Engineers, [illegible] ional Group on [illegible] 193-197.

95. Stumpers, F. L., "A Bibli[illegible] of Information [illegible] Cybernetics," Transac[illegible] of the Institute o[illegible] ional Group on Inform[illegible] on Theory, 2, 1953.

96. _____, A Bibli[illegible]y of [illegible] tion Theory, Res[illegible] Technical Report, M[illegible], 1953.

97. Thorndike, E. L. and Lorge [illegible], The Teacher's Wo[illegible] Bureau of Publicatio[illegible] Teachers College, [illegible] 1944.

98. Transactions of the Inst[illegible] of Radio Engineers, Information Theory, I[illegible] ee Jackson [12]), IT-[illegible] 1953; 3, 1954.

99. Wiener, Norbert, Cybernetic[illegible] John Wiley and Sons.

100. _____, Extrapolation, Inte[illegible] tion, and Smoothin[illegible] John Wiley and Sons.

101. Wilks, S. S., "The Likeli[illegible] Test of Independence Annals of Mathematic[illegible] tistics, 6, 1935, 1[illegible]

102. Zipf, G. K., Human Behavio[illegible] and the Principle of [illegible] Press, [illegible] 1[illegible]

UNCLASSIFIED